O'REILLY®

Аналитическая культура

ОТ СБОРА ДАННЫХ ДО БИЗНЕС-РЕЗУЛЬТАТОВ

Карл Андерсон

Эту книгу хорошо дополняют:

Маркетинг, основанный на данных
Марк Джеффри

Управление на основе данных
Тим Филлипс

О чем говорят цифры
Том Дэвенпорт и Ким Джин Хо

Большие данные
Виктор Майер-Шенбергер и Кеннет Кукьер

Верховный алгоритм
Педро Домингос

Carl Anderson

Creating a Data-Driven Organization

O'Reilly
2015

Карл Андерсон

Аналитическая культура

От сбора данных до бизнес-результатов

Перевод с английского Юлии Константиновой

Москва
«Манн, Иванов и Фербер»
2017

УДК 658.5:005
ББК 65.291.213
А65

Научный редактор Руслан Салахиев

Издано с разрешения O'Reilly Media, Inc.

На русском языке публикуется впервые

Андерсон, Карл

А65 Аналитическая культура. От сбора данных до бизнес-результатов / Карл Андерсон ; пер. с англ. Юлии Константиновой ; [науч. ред. Руслан Салахиев]. — М. : Манн, Иванов и Фербер, 2017. — 336 с.

ISBN 978-5-00100-781-4

Это практическое пошаговое руководство по внедрению в вашей организации управления на основе данных. Карл Андерсон, директор по аналитике в компании Warby Parker, провел интервью с ведущими аналитиками и учеными и собрал кейсы, которые и легли в основу данной книги. Вы узнаете, какие процессы следует ввести на всех уровнях и как именно это сделать, с какими трудностями можно столкнуться на этом пути и как их преодолеть. Автор рассказывает об аналитической цепочке ценностей, которая поможет принимать правильные решения и достигать лучших бизнес-результатов.

Книга будет интересна CEO и владельцам бизнеса, менеджерам, аналитикам.

УДК 658.5:005
ББК 65.291.213

Все права защищены.
Никакая часть данной книги не может
быть воспроизведена в какой бы то ни было форме
без письменного разрешения владельцев авторских прав.

ISBN 978-5-00100-781-4

© 2017 Mann, Ivanov and Ferber
Authorized Russian translation of the English edition of Creating a Data-Driven Organization, ISBN 9781491916919 © 2015 Carl Anderson, published by O'Reilly Media, Inc.
This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.
© Перевод на русский язык, издание на русском языке, оформление. ООО «Манн, Иванов и Фербер», 2017

Оглавление

https://t.me/it_boooks/2

Введение ... 11

Краткий обзор ... 11

Для кого эта книга? ... 14

Структура глав ... 14

Условные обозначения ... 16

Глава 1. Что значит «на основе данных»? ... 17

Сбор данных ... 17

Доступ к данным ... 18

Составление отчетности ... 21

Оповещения ... 22

От отчетов и оповещений к анализу ... 23

Критерии управления на основе данных ... 26

Зрелость аналитических данных ... 29

Краткий обзор ... 35

Глава 2. Качество данных ... 37

Аспекты качества данных ... 38

Происхождение данных ... 57

Качество данных как совместная ответственность ... 58

Глава 3. Сбор данных ... 62

Собирайте все что можно ... 62

Расстановка приоритетов при выборе источников данных ... 66

Установление взаимосвязи ... 69

Сбор данных ... 71

Покупка данных ... 74

Сколько стоит набор данных? 78
Хранение данных 81

Глава 4. Специалисты по аналитике 83
Типы специалистов по аналитике 83
Аналитика — это командный спорт 91
Навыки и качества 95
Еще один инструмент 98
Организация работы аналитиков в компании 104

Глава 5. Анализ данных 111
Что такое анализ данных? 112
Виды анализа данных 114

Глава 6. Разработка показателей 140
Разработка показателей 141
Ключевые показатели эффективности 149

Глава 7. Сторителлинг на основе данных 157
Сторителлинг 158
Первые шаги 161
Визуализация данных 165
Подача данных 176
Основные выводы 184

Глава 8. A/B-тестирование 187
Почему A/B-тестирование? 191
Практические рекомендации по A/B-тестированию 192
Другие подходы 205
Влияние на корпоративную культуру 210

Глава 9. Принятие решений 212
Как принимают решения? 214
Что осложняет процесс принятия решения? 219

Решения ... 231
Заключение ... 239

Глава 10. Корпоративная культура на основе данных ... 242
Открытость и доверие ... 243
Повышение квалификации в области работы с данными ... 247
Сначала цели ... 250
Задавайте вопросы ... 251
Итерации и обучение ... 253
Как противостоять HiPPO ... 255
Руководство на основе данных ... 256

Глава 11. Топ-менеджмент компании с управлением на основе данных ... 259
Chief Data Officer ... 261
Chief Analytics Officer ... 273
Заключение ... 279

Глава 12. Вопросы конфиденциальности, этики и риска ... 282
Уважайте конфиденциальность ... 284
Практикуйте эмпатию ... 290
Качество данных ... 294
Безопасность ... 296
Обеспечение исполнения ... 299
Заключение ... 300

Заключение ... 302

Дополнительная литература ... 309
Аналитика ... 309
Анализ данных ... 309
Принятие решений ... 309
Визуализация данных ... 310
A/B-тестирование ... 310

Приложение A. О необоснованной эффективности данных: почему больше данных лучше? 311

Проблемы типа «ближайший сосед» 312

Проблемы относительной частотности 315

Проблемы оценки одномерного распределения 316

Проблемы многофакторности 317

Приложение B. Заявление о видении 319

Ценность 321

Реализация 322

Благодарности 324

Об авторе 326

Колофон 327

Введение

Краткий обзор

Эта книга посвящена двум основным вопросам:

1) что означает для компании управление на основе данных?
2) как компания может к нему прийти?

Многие компании считают, что, если они генерируют множество отчетов или у них много дашбордов, значит, они относятся к категории компаний с управлением на основе данных. Хотя эти виды деятельности и составляют часть того, чем занимается компания, обычно они ретроспективны, то есть часто лишь представляют прошлые или настоящие факты без обеспечения достаточного контекста, без объяснения причинно-следственных связей, а также без рекомендаций, какие шаги предпринять. Иными словами, они фиксируют произошедшее, но ничего не предписывают. В этом отношении их потенциал роста ограничен.

В противовес следует рассматривать типы перспективного анализа, такие как прогнозные модели, которые способствуют оптимизации расходов на рекламу, пополнению цепочки поставок или снижению оттока покупателей. Они отвечают на вопросы «кто», «что», «когда», «почему» и «где». На основе моделей люди дают рекомендации, делают прогнозы и интерпретируют полученные данные. Часто они становятся ключевыми факторами роста в организациях с управлением на основе данных. Сформулированные на основе данных выводы и рекомендации, если их правильно использовать, оказывают огромное потенциальное влияние на эффективность деятельности компании.

Однако для получения подобных выводов требуется, чтобы были собраны правильные, заслуживающие доверия данные, анализ был проведен качественно, выводы учитывались при принятии решений, а решения подразумевали конкретные действия, чтобы потенциал был полностью реализован. Уф! Я называю эту последовательность от сбора данных до конечного результата *аналитической цепочкой ценности*.

Последний шаг в этой цепочке чрезвычайно важен. Аналитику нельзя считать основанной на данных, если полученная информация не учитывается при принятии решений и не вызывает последующих действий. Если данные игнорируются, а большой босс делает что пожелает, сбор этих данных не имеет смысла. Управление на основе данных осуществляется в компании при наличии правильных процессов и корпоративной культуры, чтобы дорабатывать или стимулировать важные деловые решения с учетом проведенного анализа данных, который таким образом оказывает непосредственное влияние на развитие бизнеса.

Ключевую роль играет создание соответствующей корпоративной культуры. Это многосторонняя программа, включающая качество данных и обмен информацией, прием на работу и обучение аналитиков, коммуникацию, аналитическую организационную структуру, разработку показателей, A/B-тестирование[1], процессы принятия решений и многое другое. Эта книга поможет пролить свет на все эти понятия благодаря доступным объяснениям и наглядным примерам из целого ряда производственных отраслей. Кроме того, здесь приводятся практические советы и рекомендации от лидеров в области анализа и обработки данных. Надеюсь, эта книга вдохновит читателей на то, чтобы переориентировать свою деятельность и начать руководствоваться данными.

Более того, на протяжении всей книги подчеркивается важная роль, которая отводится самым разным специалистам в области обработки и анализа данных. Я убежден, что компанию с управлением на основе данных и соответствующую корпоративную культуру можно и нужно развивать не только сверху вниз — от руководства на места, — но и снизу вверх. Как отметил на форуме 2014 года Chief Data Officer Executive Forum руководитель направления по анализу и обработке данных

[1] Метод маркетингового исследования, суть которого заключается в том, что контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того чтобы выяснить, какие из изменений улучшают целевой показатель. *Прим. ред.*

компании Trulia Тодд Холлоуэй, «лучшие идеи подают сотрудники, наиболее тесно работающие с данными». Они не только напрямую имеют дело с источниками данных и способны оценить их качество и повлиять на него, не только понимают, как лучше всего их дополнить, но также «часто подают хорошие идеи по поводу товаров». Кроме того, они могут помочь повысить уровень знаний других сотрудников компании в этой области. Частично это происходит благодаря тому, что они развивают свои навыки и активно применяют их для качественного выполнения работы. Другая причина в том, что у них лучше развито предпринимательское мышление: они умеют задавать правильные вопросы и формулировать бизнес-проблемы, а затем убеждать в своих выводах и рекомендациях тех, от кого зависит принятие решения, предлагая им веское обоснование, какое влияние на бизнес способны оказать эти выводы и рекомендации.

А влияние и выгоды могут быть весьма заметными. Согласно результатам одного из отчетов[1], в котором контролировались и другие факторы, в компаниях с управлением на основе данных производительность была на 5–6% выше, чем в тех, что не практикуют подобное управление. К тому же в компаниях первой категории были выше показатель использования ресурсов, коэффициент рентабельности капитала и рыночная стоимость. Согласно данным другого отчета[2], возврат на каждый вложенный в проведение аналитики 1 долл. составляет 13,01 долл. Управление на основе данных окупается!

Ориентацию на использование данных можно представить в виде непрерывного процесса: компания всегда может повысить свой уровень управления на основе данных, улучшить качество собираемых данных и аналитического процесса, провести больше тестирований. Более того, всегда можно усовершенствовать качество процесса принятия решений. В этой книге мы обсудим отличительные черты эффективных компаний с управлением на основе данных. Мы остановимся на инфраструктуре, навыках, корпоративной культуре, необходимых для создания компании, где к данным относятся как к основному активу и используют их для принятия бизнес-решений. Кроме того, мы

---

[1] Brynjolfsson E., Hitt L. M. and Kim H. H. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? Social Science Research Network (2011). URL: http://ebusiness.mit.edu/research/papers/2011.12_Brynjolfsson_Hitt_Kim_Strength%20in%20Numbers_302.pdf.

[2] Nucleus Research. Analytics pays back $13.01 for every dollar spent. O204 (Boston, MA: Nucleus Research, 2014), 5. URL: http://nucleusresearch.com/research/single/analytics-pays-back-13-01-for-every-dollar-spent/.

рассмотрим некоторые примеры поведения, которое, наоборот, мешает бизнесу максимально эффективно использовать получаемые данные.

Таким образом, цель этой книги — вдохновить специалистов по анализу и обработке данных в компаниях эффективно выполнять свои функции, время от времени делать паузу, чтобы ответить на вопросы, максимально ли использует компания свои данные и можно ли делать это еще эффективнее. Еще одна цель — стимулировать обсуждение: для каких еще целей возможно применение этого ключевого ресурса. Никогда не рано думать об этом. Основатели компании и руководство высшего звена должны постараться внедрить принципы управления на основе данных на самых ранних этапах развития организации. Давайте узнаем больше о том, что эти принципы собой представляют.

Для кого эта книга?

Информация, здесь изложенная, поможет разработать программу внутренней аналитики и управлять ею: принимать решения, какие данные собирать и хранить, как их получать и интерпретировать, и самое важное — как действовать на их основе.

Неважно, единственный ли вы специалист по анализу и обработке данных в стартапе (и притом вынуждены выполнять еще с десяток других функций) или руководитель отдела с кучей подчиненных в зрелой компании. Если вы работаете с данными и стремитесь действовать быстрее, рациональнее и эффективнее, эта книга поможет создать не просто аналитическую программу, а соответствующую корпоративную культуру.

Структура глав

Структура книги соответствует этапам создания цепочки аналитической ценности. Первые главы посвящены непосредственно данным, в частности выбору правильных источников, обеспечению качества и достоверности. Следующий шаг в этой цепочке — анализ данных. Для качественного выполнения анализа, результаты которого можно будет эффективно использовать в дальнейшей работе, нужны профессионалы, владеющие определенными навыками и инструментами. Для обозначения этой группы сотрудников намеренно используется общий

термин «специалисты по аналитической работе», который объединяет сотрудников, занимающихся сбором, обработкой, анализом данных. Это сделано на основании убеждения, что любой член команды — от младшего аналитика без опыта работы до суперзвезды в области анализа данных — вносит свою лепту в общее дело. Мы подробнее остановимся на том, какими компетенциями должен обладать хороший аналитик, как можно развивать профессиональные навыки в этой области, а также на организационных аспектах — как помочь специалисту по аналитической работе стать частью команды или подразделения. Следующие главы посвящены непосредственно аналитической работе: выполнению анализа, разработке показателей, A/B-тестированию и рассказыванию истории. Затем мы перейдем к следующему этапу в цепочке аналитической ценности — принятию решений на основе результатов анализа. Мы рассмотрим, что может затруднять процесс принятия решения и как с этим бороться.

На протяжении всей книги прослеживается основная мысль: суть процесса управления компанией на основе данных не сводится к данным как таковым или к обладанию самым современным набором инструментов по работе с большими данными. Самое важное в этом — *корпоративная культура*. Культура организации — доминирующий фактор, который устанавливает ожидания относительно того, насколько демократичным будет процесс работы с данными, как эти данные станут использоваться внутри организации, какие ресурсы, в том числе образовательные, станут инвестироваться в использование данных как стратегического актива компании. По этой причине в главе, посвященной корпоративной культуре, мы объединим все уроки, извлеченные на разных этапах цепочки аналитической ценности. В одной из последних глав обсудим роль двух относительно новых позиций в высшем руководстве компаний: CDO (Chief Data Officer, директор по управлению данными) или CAO (Chief Analytics Officer, директор по аналитике). Тем не менее рядовые сотрудники тоже в значительной мере влияют на формирование корпоративной культуры организации, поэтому на протяжении книги мы будем напрямую обращаться к специалистам по работе с данными, подчеркивая, что именно они способны сделать для повышения своего влияния на эффективность деятельности компании. В компании, для которой управление на основе данных не просто модная тенденция, сотрудники на всех уровнях уделяют большое внимание качеству данных и их оптимальному использованию при принятии взвешенных решений и для повышения конкурентного преимущества компании.

Условные обозначения

В книге используются следующие условные обозначения.

Выделение курсивом

Применяется для обозначения новых терминов, адресов сайтов (URL), адресов электронной почты, имен файлов и расширений файлов.

`Моноширинный шрифт`

Применяется для обозначения программных элементов, таких как переменные, названия функций, базы данных, типы данных, переменные окружения, утверждения и ключевые слова.

`Моноширинный шрифт с полужирным выделением`

Применяется для обозначения команд или другого текста, который должен внести пользователь.

`Моноширинный шрифт с курсивом`

Применяется для обозначения текста, который нужно заменить переменными пользователя или переменными, которые определяются контекстом.

Этот элемент обозначает совет или рекомендацию.

Этот элемент обозначает общую информацию.

ГЛАВА 1

Что значит «на основе данных»?

https://t.me/it_boooks/2

Без данных вы просто еще один человек с собственным мнением.
Уильям Эдвардс Деминг[1]

Управление на основе данных подразумевает формирование инструментов, способностей и, что самое важное, *корпоративной культуры*, которая опирается на данные. В этой главе мы рассмотрим, что отличает компанию с управлением на основе данных. Начнем с базовых требований к их сбору и доступности. Затем остановимся подробнее на весьма важном отличии — подготовке отчетов и получении оповещений в противовес процессу анализа. Существует много различных типов перспективного анализа, отличающихся по степени сложности. Мы уделим некоторое время изучению этих типов с точки зрения их «уровня аналитики» и «аналитической зрелости», а также обсудим основные признаки «аналитически зрелой» организации. Какой она должна быть?

Начнем с ответа на первый вопрос: что означает для компании управление на основе данных?

Сбор данных

Давайте сразу озвучим несколько очевидных требований.

Требование № 1: в компании должен осуществляться сбор данных.

Несомненно, данные — ключевой компонент. При этом речь идет не о любых данных, а о *правильных*. Необходимо, чтобы набор данных

[1] Уильям Эдвардс Деминг (William Edwards Deming, 1900–1993) — американский ученый, статистик и консультант по менеджменту. Создатель теории менеджмента, основанной на предложенной им же теории глубинных знаний. *Прим. перев.*

соответствовал вопросу, который требуется решить. Помимо этого, данные должны быть своевременными, точными, чистыми, объективными, и, что важнее всего, они должны заслуживать доверия.

Это не так-то просто. Данные никогда не бывают настолько чистыми, как вам кажется. Они могут быть предвзятыми, что может повлиять на результат анализа, а очистка данных может стать трудоемким и дорогим процессом, требующим времени. Часто приходится слышать, что специалисты по работе с данными до 80% времени тратят на их сбор, очистку и подготовку и только 20% — на построение моделей, процесс анализа, визуализацию и формулировку заключений на основе этих данных[1]. Как показывает опыт, это вполне вероятно.

В следующей главе мы поговорим о качестве данных подробнее.

Даже если у вас есть действительно качественные данные и даже если у вас *много* качественных данных, это означает только то, что вы обладаете этими данными, но не то, что в вашей компании действует управление на основе данных. Некоторые люди, особенно специалисты организаций, предоставляющих услуги по работе с большими данными, называют большие данные практически панацеей: если собирать абсолютно всё, где-то должен попасться алмаз (или крупинки золота, или искомая иголка, или любая другая метафора) и компания станет успешной. Горькая правда в том, что одних только данных недостаточно. Небольшое количество чистой, достоверной информации может быть гораздо более ценно, чем петабайты мусора.

Доступ к данным

Требование № 2: данные должны быть общедоступными.

Наличие точных и своевременных данных по теме еще не делает управление в вашей компании управлением на основе данных. Данные также должны отвечать еще ряду требований.

Данные могут быть объединены

Их формат должен при необходимости допускать объединение с другими данными компании. Варианты могут быть разные: реляционные базы данных, хранилища NoSQL или Hadoop. Используйте инструмент, который отвечает вашим конкретным требованиям. Например, в течение длительного времени финансовые

[1] См., например: http://bit.ly/nyt-janitor и http://bit.ly/im-data-sci.

аналитики в компании Warby Parker использовали Excel для вычисления основных показателей, которые они предоставляли высшему руководству. Они собирали огромное количество сырых данных из разных источников и запускали функцию ВПР (VLOOKUP — функцию в Excel для поиска перекрестных ссылок в данных), чтобы объединить весь массив данных и взглянуть на них в перспективе. Изначально это работало, но по мере того как базы данных по клиентам и продажам быстро росли и информации становилось все больше, объем файла в Excel начал приближаться к 300 MB, загрузка оперативной памяти компьютеров была максимальной, а обработка файла с помощью функции ВПР начала занимать до десяти часов и больше, при этом программа периодически зависала, и ее приходилось запускать заново. Специалисты компании применяли этот инструмент и подход так долго, как могли, но если когда-то Excel была вполне удобным инструментом, то динамичный рост компании изменил ситуацию. Механика получения этих данных превратилась для аналитиков в «пожиратель времени» и источник стресса: они никогда не знали, получат ли необходимые им данные или через десять часов им вновь придется перезапускать функцию ВПР. Условно говоря, из специалистов по анализу данных они превратились в специалистов Microsoft по сбору данных. Моя команда помогла перенести весь массив информации в реляционную базу данных в MySQL. Мы написали запросы для обработки данных для аналитиков, чтобы они могли сосредоточиться на анализе, выявлении трендов и презентации этих данных, что было гораздо более эффективным использованием их рабочего времени. Теперь, когда в их распоряжении более эффективные инструменты и больше времени, они способны проводить более глубокий анализ.

Данные можно использовать совместно

Внутри организации следует развивать культуру обмена данными, чтобы была возможность их сопоставлять и объединять, например связать историю поисковых запросов пользователя и историю осуществленных им покупок. Представим ситуацию: пациента доставили в отделение экстренной медицинской помощи, где ему оказали первую помощь, а затем выписали, и теперь ему необходимо обратиться за амбулаторным лечением и провести обследования. Очевидно, что качество обслуживания и, что важнее, качество лечения пострадают, если между этими медицинскими учреждениями не будет организован обмен информацией:

когда и по какой причине пациент обратился за медицинской помощью, какое лечение ему было оказано и так далее. С точки зрения представителей здравоохранения, невозможно проанализировать или улучшить процесс в отсутствие связной и четкой картины потока пациентов, процесса диагностики и полных данных наблюдения за этими пациентами за длительный срок. Таким образом, разрозненные данные всегда стараются охватить все, что возможно. Когда больший объем данных доступен для большего количества частей системы, целое всегда бывает лучше суммы частей.

Доступны по запросу

Необходимы адекватные инструменты для работы с данными и предоставления информации по запросу. В процессе анализа и составления отчетности огромный объем сырых данных необходимо отфильтровать, сгруппировать и объединить в небольшие наборы высокоуровневых показателей, чтобы обеспечить понимание того, что происходит в бизнесе. Например, мне нужно увидеть тренд или понять разницу между сегментами покупателей. У специалистов по работе с данными должны быть инструменты, позволяющие сделать это относительно просто.

(Все эти аспекты мы подробнее проанализируем в следующих главах.)

Итак, теперь у нас есть данные и доступ к ним. Достаточно ли этого? Нет, пока недостаточно. Нужны квалифицированные специалисты, которые смогут работать с этими данными. И здесь важны не только механизмы сортировки и систематизации данных, например посредством языка запросов или макросов Excel, но, главным образом, специалисты, которые будут выбирать соответствующие показатели (подробнее об этом в главе 6). К этим показателям могут относиться уровень повторной подписки (для таких сервисов, как Netflix или Wall Street Journal), долгосрочные показатели ценности или показатели роста, но в любом случае кто-то должен решать, какие именно это будут показатели, и кто-то должен создать процесс их получения.

Таким образом, человеческий фактор в управлении компанией на основе данных — важнейший: необходимы люди, способные задавать правильные вопросы, люди с необходимыми навыками для получения нужных данных и показателей, люди, использующие данные для планирования следующих шагов. Иными словами, одни лишь данные мало чем помогут компании.

Составление отчетности

Предположим, у вас есть аналитическая группа с доступом к точным данным. Эта группа получает данные по объему продаж и гордо рапортует о росте портфеля заказов компании на 5,2% с апреля по май (рис. 1.1).

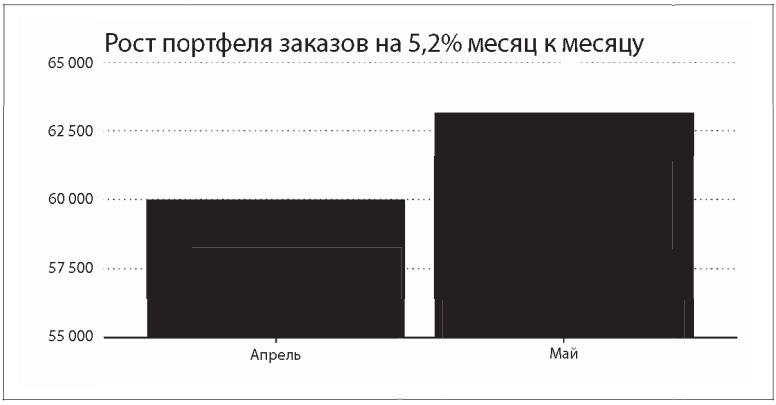


Рис. 1.1. Рост уровня продаж на 5,2% месяц к месяцу!

Кажется, что в компании осуществляется управление на основе данных. Однако этого по-прежнему недостаточно. Разумеется, хорошо, что специалисты отслеживают данные по продажам. Генерального и финансового директоров эти цифры, несомненно, заинтересуют. И тем не менее — о чем на самом деле говорит показатель 5,2%? Практически ни о чем. Возможны самые разные причины роста объема продаж компании.

— Предположим, вы продаете сезонный товар, например купальные костюмы. Может быть, рост в 5,2% — это гораздо ниже, чем обычно. Может быть, в предыдущие годы рост объема продаж в мае составлял более 7%, а в этом году он ниже обычного.

— Возможно, директор по маркетингу потратил кучу денег на национальную кампанию по повышению узнаваемости бренда. Какой процент роста из этих 5,2% обусловлен проведенной кампанией? Насколько эффективным оказалось подобное вложение средств?

— Может быть, генерального директора вашей компании пригласили поучаствовать в телешоу Good Morning America[1], или ваш

[1] Good Morning America («Доброе утро, Америка») — американское телевизионное шоу, которое транслируется по утрам на канале ABC. Выходит в эфир с 1975 г. *Прим. ред.*

продукт был упомянут в Techcrunch[1], или ваше видео стало «вирусным», и это послужило фактором роста продаж. То есть причина — какое-то конкретное событие, способное обеспечить временный или устойчивый рост.

— Возможно, продажи за месяц характеризуются низким объемом и широким ассортиментом. Возможно, это было лишь удачным стечением обстоятельств, а общая тенденция — *нисходящая*. (Если вы когда-нибудь пробовали играть на бирже, то понимаете, о чем речь.)

— Может быть, ошибка в самих данных. Если уровень продаж относительно стабилен и вы видите резкий скачок без каких-либо предпосылок к тому, возможно, все дело в качестве данных.

Все это возможные объяснения. Цифра в отчете представляет собой именно это — числовой показатель без контекста.

> «По мере того как компании становятся все более крупными и сложноорганизованными, руководство все меньше зависит от личного опыта и все больше — от обработанных данных». — Джон Гарднер
>
> Джон Маэда (@johnmaeda)
> 16 августа 2014 года[2]

Оповещения

Дзынь, дзынь, дзынь! Загрузка CPU (ЦП) на сервере приложений № 14 за последние пять минут превысила 98%.

Оповещения фактически представляют собой отчеты о том, что происходит в настоящее время. Обычно они обеспечивают конкретные данные в рамках тщательно разработанных показателей. К сожалению, как и отчеты, они не сообщают, почему наблюдается рост загрузки ЦП, и не говорят, что следует предпринять прямо сейчас для решения проблемы, то есть они не дают важного контекста.

Нет причинно-следственного объяснения. Это момент, когда системные администраторы или инженеры по эксплуатации начинают изучать журнал регистрации событий, чтобы понять, что происходит,

---

[1] Techcrunch — сайт и одноименная компания, блог, описывающий продукты, стартапы и другие сайты, основанный Майклом Аррингтоном в 2005 г. *Прим. ред.*

[2] URL: http://bit.ly/maeda-gardner.

почему и как это исправить: сделать откат назад, раскрутить дополнительные серверы, перенастроить выравниватель нагрузки и так далее.

На рис. 1.2 приведен пример загрузки сервера. С небольшими вариациями на протяжении дня очередь выполнения составляет 0,5 или меньше. В час ночи загрузка начинает расти и за 30 минут увеличивается до пяти и выше, в десять раз по сравнению с «нормой». Ситуация нестандартная. Что происходит? Возможно, требуется вмешательство? Но что нужно сделать?

В данном случае это всего лишь еженедельное резервное копирование данных. Оно осуществляется каждый четверг в час ночи. Это абсолютно штатная ситуация. Мы имеем четкие данные и ясно представленные показатели. Нет только контекста: что причина повышения загрузки — резервное копирование данных, что оно ожидаемо и запланированно происходит в определенное время и что сервер спокойно справляется с этой загрузкой.

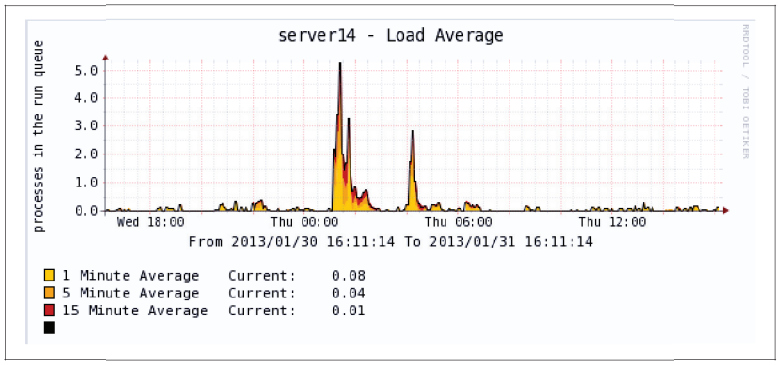


Рис. 1.2. Пример загрузки сервера

Источник: https://blog.bigwetfish.hosting/we-got-your-back/

От отчетов и оповещений к анализу

Составление отчетов и получение оповещений — необходимые факторы управления на основе данных, но этого недостаточно. Хотя не стоит недооценивать важность двух этих видов деятельности. Подготовка отчетов чрезвычайно важна для управления на основе данных: компания не сможет быть эффективной без этого элемента. А вот обратное не обязательно верно: существует множество организаций, сосредоточенных на отчетности, у которых может не быть качественного

анализа. Составление отчетности может быть вызвано официальными требованиями, например необходимостью исполнения закона Сарбейнза–Оксли[1] и подготовки отчетов о прибыли для акционеров, а не внутренним стремлением к повышению эффективности бизнеса.

Данные отчетов информируют, что произошло в прошлом. Кроме того, они могут быть тем фундаментом, с которого можно наблюдать за изменениями и тенденциями. Они могут представлять интерес для инвесторов и акционеров, но в целом это ретроспективный взгляд на ситуацию. Для управления на основе данных нужно двигаться дальше. Необходимо *прогнозировать* развитие ситуации, на основе анализа стараться понять, почему меняются показатели, и, где возможно, проводить эксперименты для сбора данных, которые могут помочь понять причины.

Давайте сравним два этих понятия. Вот варианты их возможных определений.

Отчетность — процесс организации данных в информационные сводки для отслеживания того, как функционируют разные сферы бизнеса[2].

Анализ — преобразование данных в выводы, на основе которых будут приниматься решения и осуществляться действия с помощью людей, процессов и технологий[3].

Отчет показывает, что произошло: в четверг в 10:03 на сайте наблюдалось максимальное число посетителей — 63 000 человек. Он дает конкретные цифры.

Анализ показывает, почему это произошло: в 10:01 о компании упомянули в ТВ-шоу 60 Minutes, — и рекомендует, что компании следует делать, чтобы оставаться примерно на этом же уровне.

Отчеты ретроспективны, анализ дает рекомендации.

В табл. 1.1 суммированы отличия между этими понятиями. Теперь должно быть очевидно, почему анализ и управление на основе данных — настолько важный компонент ведения бизнеса. Это факторы,

---

[1] Закон от 30 июля 2002 года, названный по именам его разработчиков и инициаторов: сенатора-демократа Пола Сарбейнза и конгрессмена-республиканца Майка Оксли. В соответствии с этим законом значительно ужесточились требования к финансовой отчетности. *Прим. ред.*

[2] Dykes B. Reporting vs. Analysis: What's the Difference? Digital Marketing Blog, October 19, 2010. URL: https://blogs.adobe.com/digitalmarketing/analytics/reporting-vs-analysis-whats-the-difference/.

[3] Faria M. Acting on Analytics: How to Build a Data-Driven Enterprise. BrightTALK, September 11, 2013. URL: https://www.brighttalk.com/webcast/1829/80223.

способные дать компании новые направления развития или вывести ее на новый уровень эффективности.

Таблица 1.1. Основные характеристики отчета и анализа

Отчет	**Анализ**
Описательный	Дает рекомендации
Что?	Почему?
Ретроспективный	Перспективный
Поднимает вопросы	Отвечает на вопросы
Данные → информация	Данные + информация → выводы
Отчеты, дашборды, оповещения	Наблюдения, рекомендации, прогнозы
Отсутствие контекста	Контекст + история

Источник: взято преимущественно у Б. Дайкса

Полезно для понимания аналитики ознакомиться с работой Т. Дэвенпорта и др. (см. табл. 1.2)[1].

Таблица 1.2. Гипотетические основные вопросы, на которые отвечает аналитика, по Дэвенпорту (на основе работы Дэвенпорта и др., 2010). Пункт D представляет собой ценную аналитику, пункты E и F обеспечивают управление на основе данных, если эта информация стимулирует конкретные действия (подробнее об этом ниже).

	Прошлое	**Настоящее**	**Будущее**
Информация	**A) Что случилось?** Отчет	**B) Что происходит сейчас?** Оповещение	**C) Что произойдет?** Экстраполяция
Выводы	**D) Как и почему это произошло?** Моделирование, экспериментальное планирование	**E) Какой следующий оптимальный шаг?** Рекомендации	**F) Что самое хорошее/плохое может произойти?** Прогноз, оптимизация, симуляция

В нижнем ряду таблицы отражены действия, приводящие к выводам. Как уже отмечалось ранее, составление отчетов (A) и оповещение (B) — не управление на основе данных: они отмечают, что уже произошло или что необычное или нежелательное происходит сейчас,

[1] Davenport T. H., Harris J. G. and Morison R. Competing on Analytics. Boston: Harvard Business Press, 2010.

но при этом не дают объяснений, почему это произошло или происходит, и не дают рекомендаций по улучшению ситуации. Предвестником управления на основе данных служит дальнейшее изучение причинно-следственных связей с помощью моделей или экспериментов (D). Только понимая причины произошедшего, можно сформулировать план действий или рекомендации (E). Пункты E и F обеспечивают управление на основе данных, но только если полученная информация стимулирует конкретные действия.

(Пункт C представляет собой опасную зону, поскольку слишком велик соблазн распространить существующий тренд на будущее: в Excel выберите «Диаграмма» (Chart), нажмите «Добавить линию тренда» (Add trendline) — и вот вы уже экстраполировали текущие данные на другие ячейки и делаете необоснованные прогнозы. Даже при обдуманном выборе функциональной формы модели может быть множество причин, почему этот прогноз ошибочен. Для уверенности в прогнозах следует использовать модель учета причинно-следственных связей. Подробнее об этом типе анализа — в главе 5.)

Итак, в нижнем ряду таблицы отражены перспективные виды деятельности, включающие элементы причинно-следственного объяснения. Теперь мы переходим к тому, что означает управление на основе данных.

Критерии управления на основе данных

Для компаний с управлением на основе данных характерны виды деятельности, перечисленные ниже.

— Эти компании постоянно проводят различные тестирования, например A/B-тестирование на сайте или тестирование заголовков в электронной рассылке маркетинговой кампании. Социальная сеть LinkedIn, например, проводит до 200 тестирований в день, сайт электронной коммерции Etsy одновременно может проводить до десяти тестирований. Тестирование иногда проводится непосредственно с участием конечных пользователей, чтобы компания могла получить прямую обратную связь относительно потенциальных новых характеристик или новых продуктов.

— Тестирования направлены на постоянное совершенствование деятельности компании и ее сотрудников. Это может быть постоянная оптимизация основных процессов, например сокращение производственного процесса на несколько минут или снижение

цены за конверсию, что становится возможным благодаря тщательному анализу, специально разработанным математическим или статистическим моделям и симуляции.

— Компании могут заниматься прогнозным моделированием, прогнозированием объема продаж, курса акций или выручки, но, что самое важное, они используют собственные прогнозные ошибки для улучшения своих моделей (см. главу 10).

— Практически всегда они выбирают среди будущих вариантов или действий на основе набора взвешенных показателей.

Ресурсы всегда конечны, и всегда есть аргументы за и против разных рациональных способов действий. Для принятия окончательного решения необходимо собрать данные для каждого набора показателей, которые тревожат или интересуют компанию, и определить их значимость. Например, когда компания Warby Parker собиралась открывать первый офис за пределами Нью-Йорка, то комплексно рассматривала и оценивала целый ряд переменных в отношении нового места: индекс благополучия Gallup (Well-being index), кадровый потенциал, прожиточный уровень, стоимость билетов до Нью-Йорка и так далее. Марисса Майер (CEO компании Yahoo!) делилась похожей историей: как она выбирала между разными предложениями о работе и приняла решение работать в компании Google[1].

Компания с управлением на основе данных будет делать хотя бы что-то из перечисленного, что направлено на будущее и имеет акцент на данных.

Итак, у нас в компании есть качественные данные и квалифицированные специалисты по работе с этими данными, которые занимаются деятельностью, направленной на перспективу. Теперь-то нас можно назвать компанией с управлением на основе данных?

К сожалению, не совсем. Это все равно что в лесу падает дерево, но никто этого не слышит. Если специалисты по работе с данными проводят анализ, но никто не обращает на него внимания, и если результаты этого анализа никак не отражаются на процессе принятия решений в компании, то это нельзя считать управлением на основе данных. Специалисты по работе с данными должны информировать тех, кто при-

[1] Bosker B. Google Exec Marissa Mayer Explains Why There Aren't More Girl Geeks. The Huffington Post, July 6, 2011. URL: http://www.huffingtonpost.com/2011/07/06/google-marissa-mayer-women-in-tech_n_891167.html.

нимает решения, и последние должны делать это, учитывая результаты работы аналитиков.

Дайкс предлагает термин «аналитическая цепочка ценности» (см. табл. 1.3). Данные ложатся в основу отчетов, которые будут способствовать проведению более глубокого анализа. Результаты анализа предоставляются лицам, принимающим решения, и процесс принятия решений строится на их основе. Это ключевой шаг. Данные и результаты анализа, о которых идет речь, требуются для принятия решения, способного повлиять на стратегию или тактику компании или ее развитие.

Технологии и обучение могут обеспечить первую часть плана: помочь специалистам по работе с данными с проведением анализа и представить результаты этого анализа. Однако именно от *корпоративной культуры* компании зависит, обратят ли на данные и результаты анализа внимание, будут ли им доверять и предпринимать на их основе конкретные действия.

Наконец мы добрались до самого важного аспекта, определяющего управление на основе данных. Для компании с управлением на основе данных именно данные — основной фактор, обусловливающий стратегию и влияющий на нее. В такой компании формируется конструктивная корпоративная культура, при которой данным доверяют, а результаты анализа бывают высокозначимыми, информативными и используются для определения следующих шагов.

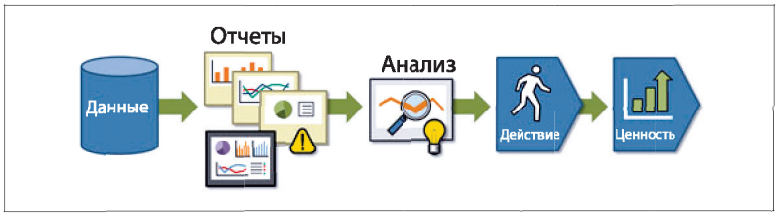


Рис. 1.3. Аналитическая цепочка ценности (по Дайксу, 2010). В компании с управлением на основе данных данные ложатся в основу отчетов, способствующих проведению более глубокого анализа. Результаты анализа влияют на процесс принятия решений, определяющий направление, в котором движется компания, и обеспечивающий ценность

Источник: https://blogs.adobe.com/digitalmarketing/analytics/reporting-vs-analysis-whats-the-difference/

В этом-то и заключается сложность. Если решения в компании принимаются на основе интуиции, как вывести ее на уровень управления на основе данных? Это процесс нелегкий и небыстрый, поэтому не стоит ожидать мгновенных изменений, однако все сотрудники компании могут внести свой вклад в этот процесс. Мы рассмотрим несколько способов, как стимулировать развитие в компании управления на основе данных.

Зрелость аналитических данных

В 2009 году Джим Дэвис, старший вице-президент и директор по маркетингу SAS Institute, выделил восемь уровней аналитических данных[1].

Стандартные отчеты

Что произошло? Когда произошло? Например, ежемесячные финансовые отчеты.

Ad hoc[2] отчеты

Как много? Как часто? Например, специальные отчеты.

Детализация по запросу
(или интерактивная аналитическая обработка, OLAP)

В чем конкретно проблема? Как найти ответы? Например, исследование данных о типах сотовых телефонов и поведении их пользователей.

Оповещения

Когда нужно действовать? Какие действия нужно предпринять немедленно? Например, загрузка ЦП, о которой говорилось ранее.

Статистический анализ

Почему это происходит? Какие возможности я упускаю? Например, почему все больше клиентов банков перекредитовываются для выплаты ипотеки.

---

[1] SAS, Eight Levels of Analytics (Cary, NC: SAS Institute, Inc., 2008), 4. URL: https://www.sas.com/en_us/news.htmlsascom/analytics_levels.pdf.

[2] Латинская фраза, означающая «к этому, для данного случая, для этой цели». В данном контексте — специальные отчеты для исследования какой-то конкретной темы. *Прим. науч. ред.*

Прогнозирование

Что, если этот тренд продолжится? Какой объем потребуется? Когда он потребуется? Например, компании, работающие в розничной торговле, могут прогнозировать спрос на продукты в зависимости от магазина.

Прогнозное моделирование

Что произойдет дальше? Как это повлияет на бизнес? Например, казино прогнозируют, кто из VIP-посетителей будет больше заинтересован в конкретных пакетных предложениях по отдыху.

Оптимизация

Как улучшить наши процессы? Какое решение сложной проблемы будет самым эффективным? Например, каков лучший способ оптимизировать ИТ-инфраструктуру с учетом многочисленных конфликтующих ограничений с точки зрения бизнеса и ресурсов?

Представленные идеи формируют график из книги Дэвенпорта и Харриса Competing on Analytics (2006)[1, 2], как показано на рис. 1.4.

(Как видите, табл. 1.2 основана на этом графике. Можно соотнести первые четыре уровня графика с верхним рядом таблицы, а вторые четыре — с нижним рядом.)

Мне нравится общая концепция и названия. Однако, исходя из того, как Дэвис (2009) и Дэвенпорт и Харрис (2007) представили свои идеи, особенно с большой восходящей стрелой, можно интерпретировать эти уровни как последовательность, своего рода иерархию, где подняться на следующий уровень можно только при условии прохождения предыдущего.

Эту псевдопрогрессию часто называют зрелостью аналитических данных. Если забьете в поисковую строку Google ключевые слова «analytics maturity», то поймете, что я имею в виду. Многочисленные специалисты представляют этот график как набор последовательных шагов для достижения цели, где односторонние стрелки указывают переход на новый уровень.

---

[1] Издана на русском языке: Дэвенпорт Т., Харрис Д. Аналитика как конкурентное преимущество. Новая наука побеждать. М. : BestBusinessBooks, 2010. *Прим. ред.*

[2] Несмотря на то что книга Дэвенпорта и Харриса появилась на два года раньше, этот источник называют «адаптация графика, сделанного компанией SAS».

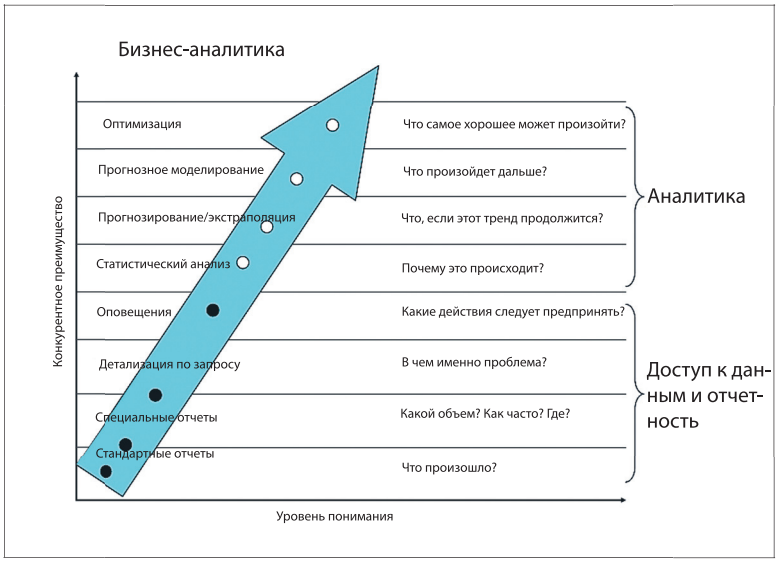


Рис. 1.4. «Бизнес-информация и аналитика» из книги Дэвенпорта и Харриса Competing on Analytics

Источник: HBR Press, ранее взято из уровней аналитических данных Джима Дэвиса

Аналитическая работа отличается от этого представления: в одно и то же время разные подразделения компании могут проводить анализ разной степени сложности.

Рон Шевлин рационально отмечает[1]:

> С точки зрения возможностей нет причин, почему компания не может прогнозировать, например, объем продаж («уровень» 6), не зная, в чем конкретно «проблема» с продажами («уровень» 3)… Но как я, будучи руководителем, должен отвечать на вопрос «Какие действия нужно предпринять немедленно?» без понимания «Что будет, если этот тренд продолжится?» и «Что произойдет дальше?» («уровни» 6 и 7)?

Мне кажется, верный способ интерпретации — подумать о том, что *максимальный* уровень развития аналитики в компании положительно

[1] Shevlin R. The Eight Levels Of Analytics? The Financial Brand, October 27, 2009. URL: http://thefinancialbrand.com/46761/the-eight-levels-of-analytics/.

коррелирует с уровнем инвестиций в аналитику, использованием данных и прочими составляющими аналитической конкурентоспособности, о которой говорят Дэвенпорт и Харрис. Например, если аналитическая команда состоит из кандидатов и докторов наук, перед которыми поставлена задача оптимизировать глобальную цепочку сбыта, очевидно, что компания серьезно инвестирует в направление работы с данными. Если в компании принято работать только с оповещениями и специальными отчетами, значит, она в меньшей степени инвестирует в аналитическое направление и для нее в меньшей степени характерно управление на основе данных.

Можно предположить, что более сложная аналитика по умолчанию лучше и что она способна сделать компанию более конкурентоспособной. Так ли это на самом деле? В интереснейшем исследовании[1], проведенном MIT Sloan Management Review совместно с IBM Institute for Business Value, были опрошены 3 тыс. руководителей и специалистов по работе с данными в 30 отраслях: как они используют аналитическую работу и что думают о ее ценности?

Один из вопросов касался конкурентного положения компании на рынке, и для него были предложены четыре ответа:

1) значительно лучше, чем у других компаний отрасли;
2) несколько лучше, чем у других компаний отрасли;
3) наравне с другими компаниями;
4) несколько или значительно хуже, чем у других компаний отрасли.

Компании, выбравшие первый и четвертый варианты ответов, считались лидерами и аутсайдерами отрасли соответственно. Что интересно, от аутсайдеров компании-лидеры отличались следующим:

— в пять раз чаще использовали аналитику;
— в три раза чаще использовали *продвинутую аналитику*;
— в два раза чаще использовали аналитику для управления своей операционной деятельностью;
— в два раза чаще использовали аналитику для составления стратегий будущего развития.

[1] LaValle S., Hopkins M. S., Lesser E., Shockley R., Kruschwitz N. Analytics: The New Path to Value. MIT Sloan Management Review, October 24, 2010. URL: http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Несомненно, есть факторы, осложняющие эту методологию. Во-первых, так называемая ошибка выжившего[1]. Во-вторых, корреляция между успешностью компании и ее размером (насколько известно, выручка компаний, участвовавших в опросе, была в диапазоне от менее 500 млн до более чем 10 млрд долл.). Возможно, только у более крупных и более успешных организаций имелось достаточно ресурсов на создание и обеспечение функций аналитических отделов, способных на разработку моделей для имитационного моделирования цепочки поставок. Тем не менее все пришли к единому мнению, что более качественная и глубокая аналитика повышает ценность бизнеса.

Авторы исследования выделили три уровня аналитических возможностей: желательный, опытный, преобразованный. Их краткие характеристики приведены в табл. 1.3.

От организаций, находящихся на желательном уровне, организации, находящиеся на преобразованном уровне, отличаются тем, что в них:

— в четыре раза выше вероятность качественного отбора информации;

— в девять раз выше вероятность качественной обработки информации;

— в восемь раз выше вероятность качественного анализа;

— в десять раз выше вероятность качественного распространения информации;

— на 63% чаще используют централизованные аналитические отделы в качестве основного источника аналитических данных (об аналитических организационных структурах речь пойдет в главе 4).

Конечно, в этом случае также наблюдается сложное взаимодействие между причинами и следствием, но взаимосвязь между конкурентным положением компании на рынке относительно других игроков и уровнем аналитической работы, проводящейся в ней, очевидна.

[1] Систематическая ошибка выжившего (*англ.* survivorship bias) — разновидность систематической ошибки отбора, когда по одной группе («выжившим») есть много данных, а по другой («погибшим») — практически нет. Так как исследователи пытаются искать общие черты среди «выживших», то упускают из виду, что не менее важная информация скрывается среди «погибших». *Прим. перев.*

Таблица 1.3. Уровни аналитических возможностей: желательный, опытный, преобразованный

	Желательный	**Опытный**	**Преобразованный**
Используют аналитические данные для…	Оправдания действий	Руководства действиями	Планирования действий
Применяют строгие, тщательные подходы для принятия решений	Редко	Иногда	Преимущественно
Способность собирать, обрабатывать и анализировать данные или делиться информацией и выводами	Ограниченная	Средняя	Высокая
Области использования	— Финансы и бюджетирование — Операционная деятельность и производство — Продажи и маркетинг	— Все функции из колонки «Желательный» — Стратегия/развитие бизнеса — Обслуживание клиентов — Исследования и разработка новых продуктов	— Все функции из колонок «Желательный» и «Опытный» — Управление рисками — Удовлетворенность клиента качеством обслуживания — Использование персонала — Общее управление — Управление брендом и маркетингом

Источник: взято и изменено: http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

Так что же тогда мешает компаниям активно применять аналитические инструменты? Два из трех наиболее распространенных ответов на этот вопрос — недостаток понимания, как использовать аналитические данные, и недостаток навыков аналитической работы внутри компании (см. рис. 1.5).

В этих ответах перечислены причины, с которыми может справиться любой специалист-аналитик. Например, аналитики могут помочь сотрудникам «прокачать» необходимые навыки, и они сами могут более активно доносить ценность аналитической работы до руководителей. Они могут проводить больше исследований и приводить практические примеры, как другим компаниям удалось справиться

с похожими трудностями в бизнесе при помощи аналитики. Руководители специалистов по сбору и обработке данных могут выделить ресурсы на улучшение качества данных, чтобы они ни у кого не вызывали сомнения. Руководители высшего звена могут стимулировать увеличение обмена данными внутри компании, а также отдельно назначить человека, отвечающего за это направление, например CAO или CDO (подробнее об этом в главе 11). В этом процессе каждый играет свою роль.

Рис. 1.5. Ответы на вопрос «Что становится основным препятствием для активного использования информации и аналитических данных в вашей компании?»

Краткий обзор

На всех этих аспектах мы остановимся подробнее в следующих главах. Во-первых, мы изучим сырые и агрегированные данные и их качество (главы 2 и 3). Затем перейдем к аналитическим структурам: какими могут быть специалисты по аналитической работе, какими навыками они должны обладать, как должен быть организован аналитический отдел (глава 4). Мы остановимся на аспектах анализа данных (глава 5), разработки показателей (глава 6) и рассказывании историй с помощью данных (глава 7). В главе 8 речь пойдет о A/B-тестировании. Мы поговорим о корпоративной культуре и процессе принятия решений, которые представляют собой важные признаки компании с управлением

на основе данных (главы 9 и 10). Мы покажем, что изменения в корпоративной культуре и оперативном управлении возможны только благодаря руководителям, которые используют в своей работе принципы управления на основе данных. В частности, мы поговорим о трех новых управленческих позициях: CDO, Chief Digital Officer[1] (директор по цифровым технологиям) и CAO (глава 11). Глава 12 будет посвящена вопросам этики и тому, как компания, уважающая персональные данные, может ограничить их использование. В конце мы дадим общее заключение.

[1] Эту позицию принято обозначать аббревиатурой CDO, но мы будем давать ее полностью во избежание путаницы. Аббревиатуру CDO будем использовать для позиции Chief Data Officer. *Прим. ред.*

ГЛАВА 2
Качество данных

https://t.me/it_boooks/2

80% времени я трачу на очистку данных. Качественные данные всегда выигрывают у качественных моделей.

Томсон Нгуен[1]

Данные — это фундамент, на котором держится компания с управлением на основе данных.

Если люди, принимающие решения, не располагают своевременной, релевантной и достоверной информацией, у них не остается другого выхода, как только положиться на собственную интуицию. Качество данных — ключевой аспект.

В этой главе понятие «качество» употребляется в самом широком смысле и рассматривается преимущественно с точки зрения аналитической работы.

Специалистам-аналитикам нужны правильные данные, собранные правильным образом и в правильной форме, в правильном месте, в правильное время. (Они просят совсем не много.) Если какое-то из этих требований не выполнено или выполнено недостаточно хорошо, у аналитиков сужается круг вопросов, на которые они способны дать ответ, а также снижается качество выводов, которые они могут сделать на основании данных.

[1] Томсон Нгуен (Thomson Nguyen) — основатель и CEO (высшая исполнительная должность в компании; в российской иерархии аналог генерального директора) компании Framed Data, которая занимается различными проблемами данных в аналитике, инфраструктуре и машинном обучении для бизнеса и некоммерческих организаций. *Прим. перев.*

Эта и следующая главы посвящены обширной теме качества данных. Во-первых, мы обсудим, как обеспечить правильность процесса сбора данных. С этой точки зрения качество данных выражается в их точности, своевременности, взаимосвязанности и так далее. Затем, в следующей главе, мы поговорим о том, как убедиться, что мы собираем правильные данные. С этой точки зрения качество выражается в выборе оптимальных источников данных, чтобы обеспечить максимально эффективные выводы. Иными словами, мы начнем с того, как правильно собирать данные, и перейдем к тому, как собирать правильные данные.

В этой главе мы сосредоточимся на способах определения достоверности данных и рассмотрим случаи, когда данные могут оказаться ненадежными. Для начала разберем критерии качества — все характеристики чистых данных. Затем рассмотрим самые разные факторы, влияющие на ухудшение качества. Этой теме мы уделим особое внимание по ряду причин. Во-первых, подобных факторов может быть великое множество, и они носят практический, а не теоретический характер. Если вам доводилось работать с данными, то, скорее всего, вы сталкивались с большинством из них. Они неотъемлемая часть нашей реальности и возникают гораздо чаще, чем нам бы того хотелось. Именно поэтому у большинства специалистов по работе с данными подавляющая часть рабочего времени уходит на очистку. Более того, вероятность возникновения этих факторов повышается с увеличением объема данных. Мой бывший коллега Самер Масри однажды заметил: «При работе с большими масштабами данных всегда помните, что вещи, которые случаются “один раз на миллион”, могут произойти в каждую секунду!» Во-вторых (и, возможно, это даже важнее), активная проверка и сохранение качества данных — совместная обязанность всех сотрудников. Каждый участник аналитической цепочки ценности должен следить за качеством данных. Таким образом, каждому участнику будет полезно на более глубоком уровне разбираться в этом вопросе.

Итак, учитывая все сказанное, давайте рассмотрим, что означает качество данных.

Аспекты качества данных

Качество данных невозможно свести к одной цифре. Качество — это не 5 или 32. Причина в том, что это понятие охватывает целый ряд аспектов, или направлений. Соответственно, начинают выделять уровни

качества, при которых одни аспекты оказываются более серьезными, чем другие. Важность этих аспектов зависит от *контекста* анализа, который должен быть выполнен с этими данными. Например, если в базе данных с адресами клиентов везде указаны коды штатов, но иногда пропущены почтовые индексы, то отсутствие данных по почтовым индексам может стать серьезной проблемой, если вы планировали построить анализ на основе показателя почтового индекса, но никак не повлияет на анализ, если вы решили проводить его на уровне показателя по штатам.

Итак, качество данных определяется несколькими аспектами. Данные должны отвечать ряду требований.

Доступность

У аналитика должен быть доступ к данным. Это предполагает не только разрешение на их получение, но также наличие соответствующих инструментов, обеспечивающих возможность их использовать и анализировать. Например, в файле дампа памяти SQL (Structured Query Language — языка структурированных запросов при работе с базой данных) содержится информация, которая может потребоваться аналитику, но не в той форме, в которой он сможет ее использовать. Для работы с этими данными они должны быть представлены в работающей базе данных или в инструментах бизнес-аналитики (подключенных к этой базе данных).

Точность

Данные должны отражать истинные значения или положение дел. Например, показания неправильно настроенного термометра, ошибка в дате рождения или устаревший адрес — это все примеры неточных данных.

Взаимосвязанность

Должна быть возможность точно связать одни данные с другими. Например, заказ клиента должен быть связан с информацией о нем самом, с товаром или товарами из заказа, с платежной информацией и информацией об адресе доставки. Этот набор данных обеспечивает полную картину заказа клиента. Взаимосвязь обеспечивается набором идентификационных кодов или ключей, связывающих воедино информацию из разных частей базы данных.

Полнота

Под неполными данными может подразумеваться как отсутствие части информации (например, в сведениях о клиенте не указано его имя), так и полное отсутствие единицы информации (например, в результате ошибки при сохранении в базу данных потерялась вся информация о клиенте).

Непротиворечивость

Данные должны быть согласованными. Например, адрес конкретного клиента в одной базе данных должен совпадать с адресом этого же клиента в другой базе. При наличии разногласий один из источников следует считать основным или вообще не использовать сомнительные данные до устранения причины разногласий.

Однозначность

Каждое поле, содержащее индивидуальные данные, имеет определенное, недвусмысленное значение. Четко названные поля в совокупности со словарем базы данных (подробнее об этом чуть позже) помогают обеспечить качество данных.

Релевантность

Данные зависят от характера анализа. Например, исторический экскурс по биржевым ценам Американской ассоциации землевладельцев может быть интересным, но при этом не иметь никакого отношения к анализу фьючерсных контрактов на грудинную свинину.

Надежность

Данные должны быть одновременно полными (то есть содержать все сведения, которые вы ожидали получить) и точными (то есть отражать достоверную информацию).

Своевременность

Между сбором данных и их доступностью для использования в аналитической работе всегда проходит время. На практике это означает, что аналитики получают данные как раз вовремя, чтобы завершить анализ к необходимому сроку. Недавно мне довелось узнать об одной крупной корпорации, у которой время ожидания

при работе с хранилищем данных составляет до одного месяца. При такой задержке данные становятся практически бесполезными (при сохранении издержек на их хранение и обработку), их можно использовать только в целях долгосрочного стратегического планирования и прогнозирования.

Ошибка всего в *одном* из этих аспектов может привести к тому, что данные окажутся частично или полностью непригодными к использованию или, хуже того, будут казаться достоверными, но приведут к неправильным выводам.

Далее мы остановимся на процессах и проблемах, способных ухудшить качество данных, на некоторых подходах для определения и решения этих вопросов, а также поговорим о том, кто отвечает за качество данных.

ДАННЫЕ С ОШИБКАМИ

Ошибки могут появиться в данных по многим причинам и на любом этапе сбора информации. Давайте проследим весь жизненный цикл данных с момента их генерации и до момента анализа и посмотрим, как на каждом из этапов в данные могут закрадываться ошибки.

В данных всегда больше ошибок, чем кажется. По результатам одного из исследований[1], ежегодно американские компании терпят ущерб почти в 600 млн долл. из-за ошибочных данных или данных плохого качества (это 3,5% ВВП!).

Во многих случаях аналитики лишены возможности контролировать сбор и первичную обработку данных. Обычно они бывают одним из последних звеньев в длинной цепочке по генерации данных, их фиксированию, передаче, обработке и объединению. Тем не менее важно понимать, какие проблемы с качеством данных могут возникнуть и как их потенциально можно разрешить.

Цель этой части книги — выделить общие проблемы с качеством данных и возможные подводные камни, показать, как избежать этих проблем и как понять, что эти проблемы присутствуют в наборе данных. Более того, чуть позже вы поймете, что это призыв ко всем специалистам,

[1] Eckerson W. Data Warehousing Special Report: Data Quality and the Bottom Line (Chatsworth, CA: 101communications LLC, 2002), 34. URL: http://download.101com.com/pub/tdwi/Files/DQReport.pdf.

работающим с данными, по возможности активно участвовать в проверке качества данных.

Итак, начнем с самого начала — с источника данных. Почему в данные могут закрасться ошибки и как с этим бороться?

ГЕНЕРАЦИЯ ДАННЫХ

Генерация данных — самый очевидный источник возможных ошибок, которые могут появиться в результате технологического (приборы), программного (сбои) или человеческого факторов.

В случае технологического фактора приборы могут быть настроены неправильно, что может сказаться на полученных данных. Например, термометр показывает 35 °C вместо 33 °C на самом деле. Это легко исправить: прибор или датчик можно настроить по другому, «эталонному», прибору, отражающему достоверные данные.

Иногда приборы бывают ненадежными. Мне довелось работать в грантовом проекте Агентства передовых оборонных исследовательских проектов Министерства обороны США (DARPA), посвященном групповой робототехнике. В нашем распоряжении была группа простейших роботов, задача которых заключалась в совместном картографировании местности. Сложность состояла в том, что инфракрасные датчики, установленные на роботах, были очень плохого качества. Вместо того чтобы сосредоточиться на разработке децентрализованного алгоритма для нанесения здания на карту, бóльшую часть времени я потратил на работу с алгоритмическими фильтрами, пытаясь справиться с качеством информации от этих датчиков, измерявших расстояние до ближайшей стены или до других роботов. Значения сбрасывались, или показатель расстояния до ближайшей стены мог неожиданно измениться на целый метр (неточность > 50%), притом что робот оставался неподвижным. Информации от этих датчиков просто нельзя было верить.

Когда в сборе данных принимают участие люди, ошибки в данных могут появиться по самым разным причинам. Сотрудники могут не знать, как правильно пользоваться оборудованием, они могут торопиться или быть невнимательными, они могут неправильно понять инструкции или не следовать им. Например, в двух больницах могут по-разному измерять вес пациентов: в обуви и без обуви. Для исправления ошибок такого рода требуются четкие инструкции и обучение персонала. Как с любым экспериментом, необходимо попытаться контролировать

и стандартизировать как можно больше этапов процесса, чтобы данные оставались максимально достоверными, сравнимыми и удобными в использовании.

ВВОД ДАННЫХ

Когда данные генерируются вручную, например при измерении веса пациентов, их необходимо зафиксировать. Несмотря на обещания электронного офиса, большой объем данных сегодня по-прежнему сначала попадает на бумагу в качестве промежуточного шага до попадания в компьютер. На этом этапе может возникнуть множество ошибок.

Ошибки случаются при расшифровке документов, заполненных от руки. (Если бы вы видели мой почерк, у вас бы не осталось в этом сомнений.) Больше всего исследований в этой области проведено в сфере здравоохранения, частично потому что последствия использования неточной информации могут быть слишком серьезными, как с точки зрения здоровья пациентов, так и с точки зрения стоимости проведения ненужных медицинских тестов. Согласно результатам одного из исследований, 46% медицинских ошибок (при базовом уровне 11% от всех записей) обусловлено неточностью при расшифровке[1]. Уровень ошибок в базах данных некоторых клинических исследований достигал 27%[2]. Подобные ошибки могли быть результатом того, что медицинский персонал неправильно читал или понимал написанное от руки, не слышал или не понимал информацию из-за плохого качества аудиоисточника или непривычных слов или неправильно вносил информацию в компьютер.

Например, я работал в одной из компаний в сфере здравоохранения, и основными базами данных, которые компания использовала чаще всего, были данные статистических опросов населения в рамках Национальной программы проверки здоровья и питания (NHANES). Мобильные клиники по всей стране проводили опросы населения: измеряли вес и артериальное давление, выясняли, есть ли в семье больные

[1] Seely C. E., Nicewander D., Page R. and Dysert P. A. A baseline study of medication error rates at Baylor University Medical Center in preparation for implementation of a computerized physician order entry system. Proc (Bayl Univ Med Cent). 2004 Jul 17(3): 357–361. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200672/.

[2] Goldberg S. I., Niemerko A. and Turchin A. Analysis of Data Errors in Clinical Research Databases. AMIA Annu Symp Proc. 2008: 242–246. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656002/.

диабетом или раком, и так далее. Когда мы изучили информацию о человеческом росте в одной из баз данных по этому проекту, то обнаружили целый ряд людей с показателем роста пять дюймов (примерно 12,5 см)! Эти данные вносили в базу специально обученные сотрудники, которые изо дня в день проводили опросы населения. Поскольку измерение роста — относительно простая процедура, наиболее вероятной причиной ошибки кажется некорректный ввод информации. Возможно, рост респондентов на самом деле был пять футов и пять дюймов (примерно 162 см) или шесть футов и пять дюймов (примерно 192 см). К сожалению, поскольку мы не знали этого наверняка, нам пришлось отметить эти значения как неизвестные.

К счастью, показатель роста человека пять дюймов — это настолько очевидная ошибка, что нам удалось определить ее с помощью простой гистограммы, и мы точно понимали, что это ошибка. Однако так бывает не всегда. Есть разные степени очевидности ошибки. Предположим, что при расшифровке записей, сделанных от руки, сотрудник вместо «аллергия на кошек и собак» написал: «аллергия на окшек и собак». Слова «окшек» не существует. Очевидно, что это опечатка, а смысл легко поддается восстановлению по контексту. Более сложными могут оказаться случаи, когда при перестановке букв могут образоваться другие слова, имеющие смысл. Тогда заметить ошибку сложнее. Разобраться со смыслом можно с помощью контекста, но он не всегда служит гарантией. Наконец, представьте, что местами случайно переставили не буквы, а цифры, например в числе 56,789 поменяли две последние цифры: 56,798. Заметить ошибку в этом случае будет чрезвычайно сложно или даже невозможно.

В целом ошибки при вводе информации можно свести к четырем типам.

Запись

Введенные слова или показатели не те, что были в оригинале.

Вставка

Появление дополнительного символа: 56,789 → 564,789.

Удаление

Один или несколько символов теряются: 56,789 → 56,89.

Перемена мест

Два или более символов меняются местами: 56,789 → 56,798.

В качестве отдельных категорий «Вставки» и «Удаления» можно выделить диттографию — случайное повторение символа (56,789 → 56,7789) и гаплографию — пропуск повторяющегося символа (56,779 → 56,79). Эти термины употребляют ученые, занимающиеся восстановлением поврежденных и переписанных от руки древних текстов, и обозначают разновидность проблемы с некачественными данными.

Особенно часто опечатки встречаются в написании дат. Например, я британец, и в английской культуре принят определенный формат написания даты: день/месяц/год. Однако я живу в США, где формат написания даты отличается: месяц/день/год. Первые несколько лет жизни в США я постоянно путался, и могу предположить, что эта проблема знакома не только мне. Представьте себе сайт, на котором пользователи со всего мира вводят в специальное поле дату. У пользователей из разных стран могут быть разные ожидания относительно формата ввода этой информации, и без необходимых подсказок могут возникнуть ошибки при вводе данных. Некоторые их них легко заметить: например, 25 марта (3/25 в американском варианте) — 25 явно не может быть обозначением месяца. А как насчет 4/5? Вы уверены, что для всех пользователей эта дата обозначает 5 апреля?

Как бороться с такого рода ошибками?

Снижение количества ошибок при вводе данных

Первый шаг, если он возможен, заключается в сокращении количества этапов от генерации данных до ввода. Скажу очевидное: если есть возможность избежать бумажной формы, лучше сразу вносить данные в компьютер.

Везде, где возможно, добавьте проверку значения каждого поля в свою электронную форму (рис. 2.1). То есть если данные четко структурированы и имеют установленный формат (например, почтовый индекс в США содержит от пяти до девяти цифр, а номер социальной страховки состоит из девяти цифр), проверяйте данные на соответствие этому формату, в противном случае предложите пользователю исправить возможные ошибки. Процесс проверки не ограничен только числовыми значениями. Например, можно проверять, чтобы дата или время вылета «обратно» были позже, чем вылета «туда». Иными словами, проверяйте все что можно, чтобы максимально избежать «мусора» в самом начале.

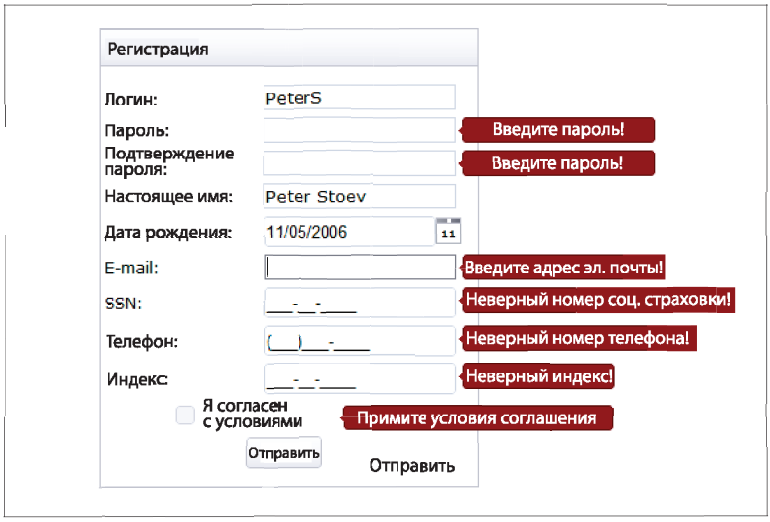


Рис. 2.1. Пример проверки значений в онлайновой регистрационной форме

Источник: http://www.jqwidgets.com

Если есть ограниченный набор допустимых значений, например аббревиатуры названий штатов в США, предложите пользователю выбрать нужный вариант из меню выпадающего списка. Автозаполнение может стать еще одним вариантом. В целом стремитесь к тому, чтобы пользователю пришлось вводить как можно меньше данных: лучше предложить варианты ответа на выбор, если, конечно, это позволяет формат требуемой информации.

В идеале постарайтесь максимально исключить человеческий фактор при сборе данных и по возможности автоматизируйте этот процесс.

Если вы располагаете временем и ресурсами, поручите двум сотрудникам независимо друг от друга расшифровывать данные (или пусть это дважды делает один сотрудник), сравнивать результаты и перепроверять данные в случае расхождений. Этот метод известен как «принцип двойной записи». Однажды я поручил стажеру расшифровать параметры из набора технических чертежей, он сделал это, а затем по собственной инициативе выполнил работу еще раз с последующей проверкой на различия. Мне как получателю данных это обеспечило уверенность в том, что точность данных максимально соответствует моим ожиданиям.

Интересный метод проверки применяется при передаче важных данных в цифровой форме, например номеров банковских счетов,

номеров социальной страховки или даже номера ISBN этой книги. Этот метод называется *контрольное число*. После передаваемого номера добавляется число, которое представляет собой определенную функцию остальных цифр номера, и это число используется для проверки того, что предыдущие цифры были переданы из системы в систему без ошибок. Предположим, вам нужно передать индекс 94121. Воспользуемся самой простой схемой. Последовательно сложим все цифры, составляющие наш индекс, и получим 17. Сложим и эти цифры, получим 8. Передаем число 941218. Принимающая система выполняет все те же самые операции, но в обратной последовательности. Она отсекает последнюю цифру: 94121 → 17 → 8. Проверяет сумму цифр и получает в итоге 8. Почтовый индекс передан верно. В случае ошибки при передаче данных, например если бы вы передали почтовый индекс 841218, система обнаружила бы ошибку при проверке: 84121 → 16 → 7 ≠ 8.

Эта схема не отличается надежностью: 93221 (случайное повторение символа) или 94211 (перестановка символов местами) эту проверку пройдут. В случае необходимости контрольного числа в реальной жизни применяются более сложные математические функции, которые способны выявить в том числе и две указанные выше ошибки. Маршрутный номер (код банка, присваиваемый Американской банковской ассоциацией) — уникальное девятизначное число, стоящее в нижней части чека перед номером счета, — один из таких примеров[1]. Контрольное число маршрутного номера — функция

$$3 \times (d_1 + d_4 + d_7) + 7 \times (d_2 + d_5 + d_8) + d_3 + d_6 + d_9 \bmod 10 = 0$$

(mod означает получение остатка от целочисленного деления. Так, 32 mod 10 = 2, поскольку 32 = 3 × 10 + 2), которая проверяется простым кодом на языке Python:

```
routing_number = "122187238"
d = [int(c) for c in routing_number]
checksum = (  # do the math!
              7 * (d[0] + d[3] + d[6]) +
              3 * (d[1] + d[4] + d[7]) +
              9 * (d[2] + d[5])
            ) % 10
print(d[8] == checksum)
```

[1] Подробную информацию о маршрутном номере можно найти по ссылке: https://en.wikipedia.org/wiki/Routing_transit_number.

Как видите, есть ряд способов, позволяющих сохранить высокое качество данных на стадии ввода информации. Но, к сожалению, и их нельзя считать абсолютно надежными. Итак, у вас в системе есть данные, которые переходят на стадию анализа. Что дальше?

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

При получении любой информации аналитику в первую очередь следует в той или иной форме провести разведочный анализ данных (глава 5) для оценки их качества. Простой способ проверки на вопиющие ошибки, как в приведенном выше примере с людьми пятидюймового роста, — сделать сводку из данных. Для каждого показателя можно составить пятичисловую сводку: два крайних значения (максимальное и минимальное значение), нижний (25-й процентиль) и верхний (75-й процентиль) квартили и медиану. Посмотрите на крайние значения. Насколько они адекватны? Они выше или ниже значений, которые вы могли бы ожидать? Пять дюймов — это очевидно слишком мало.

Вот пример того, как выглядит классификация набора данных по ирисам, представленная с помощью R — бесплатной и открытой программной среды для статистических вычислений и построения графиков, которой часто пользуются специалисты по статистике и работе с данными[1]. Американский ботаник Эдгар Андерсон собрал данные о 150 экземплярах ириса, по 50 экземпляров из трех видов, а Рональд Фишер на примере этого набора данных продемонстрировал работу созданного им метода для решения задачи классификации[2].

```
> summary (ирис)
Длина чашелистика Ширина чашелистика Длина лепестка Ширина лепестка Вид ириса
Мин. :4.300       Мин. :2.000        Мин. :1.000    Мин. :0.100     setosa :50
1-й кв. :5.100    1-й кв. :2.800     1-й кв. :1.600 1-й кв. :0.300  versicolor :50
Медиана :5.800    Медиана :3.000     Медиана :4.350 Медиана :1.300  virginica :5
Средн. :5.843     Средн. :3.057      Средн. :3.758  Средн. :1.199
3-й кв. :6.400    3-й кв. :3.300     3-й кв. :5.100 3-й кв. :1.800
Макс. :7.900      Макс. :4.400       Макс. :6.900   Макс. :2.500
```

В этом виде можно легко получить общее представление о данных (1-й кв. = 1-й квартиль, или 25-й процентиль; 3-й кв. = 75-й процентиль). Ту же самую информацию можно представить в виде коробчатой диаграммы (рис. 2.2).

---

[1] URL: https://www.r-project.org/.

[2] Подробную информацию можно найти по ссылке: https://en.wikipedia.org/wiki/Iris_flower_data_set.

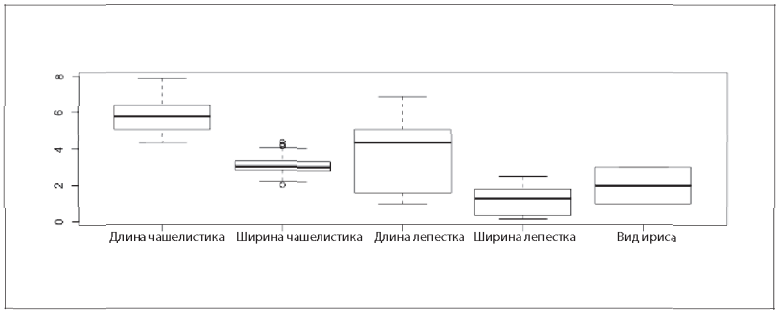


Рис. 2.2. Коробчатая диаграмма классификации набора данных по ирисам

На рис. 2.3 отражены некоторые ошибки, которые можно определить с помощью представления данных в виде простой гистограммы. В базе данных NHANES меня также интересовали данные, касающиеся артериального давления. После классификации выборки я получил максимальные значения артериального давления, которые показались мне гораздо выше нормы. Сначала я решил, что это тоже ошибка. Однако распределение показало, что эти значения хоть и находятся в хвосте распределения, но с разумной частотой. Я сверился с медицинской литературой и убедился, что значения артериального давления действительно могут быть такими высокими. Однако респондентами были люди, которые, скорее всего, не получали лечения. Как вы помните, опрос проводился среди всего населения США, а не среди пациентов медицинских учреждений, где им была бы оказана помощь, — все зависит от контекста.

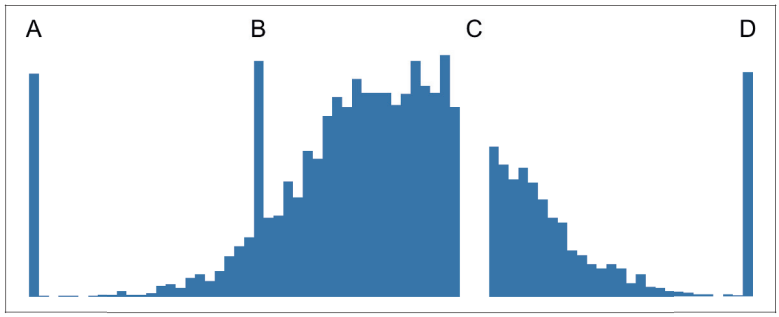


Рис. 2.3. Примеры типов ошибок, которые можно выявить с помощью простой гистограммы: A — значения по умолчанию, такие как –1, 0 или 1/1/1900; B — неправильный ввод или повтор данных; C — пропущенные данные; D — значения по умолчанию, такие как 999

Два важных навыка, которые должны развивать в себе аналитики, — прогнозирование возможных результатов и способность предварительно оценивать данные[1]. Я ошибся относительно значений артериального давления, так как оценивал их с точки зрения нормы для обычных здоровых людей. Тем не менее я узнал нечто новое для себя, скорректировал свои ожидания и убедился, что данные, скорее всего, верные.

Это наглядный пример того, что изначально вы, возможно, будете ставить под сомнение все источники данных. Я всегда исхожу из базового предположения, что данные могут быть ошибочными, и моя работа в том, чтобы выяснить источник проблемы. Я не впадаю в крайности, но непременно провожу определенную работу (например, пользуюсь функциями `summary()`, `pairs()` и `boxplot()` в R, чтобы убедиться, что в данных нет очевидных ошибок. При работе с базами данных NHANES мы с коллегами создали гистограммы всех показателей, чтобы отследить случайные образцы, бимодальное распределение и другие резко выделяющиеся значения. Подсчет числа записей на конкретную дату может послужить еще одним простым тестом. Подобный разведочный анализ данных может быть простым, быстрым и чрезвычайно ценным.

ПРОПУЩЕННЫЕ ДАННЫЕ

Одна из наиболее существенных проблем — неполные или пропущенные данные (рис. 2.3C). Эта ошибка может быть двух видов: пропуск данных в записи или пропуск всей записи.

ЗАПОЛНЯЕМ ПРОПУСКИ: МЕТОД ВОССТАНОВЛЕНИЯ

Существуют статистические подходы, которые можно применить для восстановления пропущенных данных или подстановки на их место наиболее вероятных значений (мне нравятся инструмент Amelia package от R[2] и сервис подстановки Google[3]). Их успех зависит от ряда факторов, в том числе от размера выборки, количества и характера пропущенных данных, типа переменных (являются ли они однозначными, непрерывными, дискретными и так далее), а также зашумленности данных. Один из наиболее простых подходов заключается в том, чтобы заполнить пропущенные значения средним значением

[1] Способность сделать приблизительный прогноз относится к недооцененным аналитическим навыкам. Я рекомендовал бы к прочтению главу 7 книги П. Джанерта Data Analysis with Open Source Tools (2011).

[2] URL: https://cran.r-project.org/web/packages/Amelia/index.html.

[3] URL: https://cloud.google.com/prediction/docs/smart_autofill_add_on.

этой переменной. В более сложных подходах применяются вариации EM-алгоритма[1]. Рекомендуемые к прочтению книги по этой теме: Missing Data (автор — П. Эллисон) и Statistical Analysis with Missing Data (авторы — Р. Литтл и Д. Рубин)[2]. Это эффективный инструмент, но в зависимости от типа данных сделанные с его помощью прогнозы в некоторых случаях могут быть неверными.

Зачем тогда рисковать и использовать этот подход? Во многих случаях, особенно в медицине и социальных науках, сбор данных может быть очень дорогим, к тому же возможность для сбора может быть только одна. Например, если вам нужно узнать значение артериального давления пациента на третий день клинического исследования, вы не можете вернуться в этот день, чтобы еще раз его измерить. Основная проблема заключается в том парадоксе, что чем меньше размер выборки, тем более ценна каждая запись. При этом чем меньше информации, с которой приходится работать алгоритму по восстановлению данных, тем менее точным получится результат.

Какое-то из пропущенных значений в записи способно сделать бесполезной всю эту запись. Это происходит в случае отсутствия ключевой информации, то есть показателя, определяющего тему записи (например, идентификационные данные клиента или заказа) и необходимого для объединения с другими данными. Кроме того, это может иметь место в случае, когда анализ строился на пропущенных данных. Например, если вы решили проанализировать продажи по почтовому индексу, а в какой-то записи индекс отсутствует, очевидно, что вы эту запись использовать не сможете. Если вам повезло и пропущенные данные не требуются для анализа, то выборка может и не сократиться.

Как уже говорилось ранее, причины пропуска данных могут быть самыми разными. Например, при проведении опроса респондент может не понять или пропустить вопрос, человек, обрабатывающий анкеты, может не разобрать почерк, или респондент может «на полпути» отказаться от участия в опросе. Бывает, что подводят технические средства: выходит из строя сервер или датчик. Поскольку эти причины в значительной мере влияют на качество данных, важно выяснить, *почему* данные отсутствуют.

---

[1] Expectation-maximization (EM) algorithm (*англ.*) — алгоритм, который используется в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае когда модель зависит от некоторых скрытых переменных. *Прим. науч. ред.*

[2] Последняя издана на русском языке: Литтл Р., Рубин Д. Статистический анализ данных с пропусками. М. : Финансы и статистика, 1990. *Прим. ред.*

Предположим, сломался сервер, на котором локально хранились нужные вам данные. Это может быть примером полностью потерянных записей. При наличии выравнивателя нагрузки, работающего на 20 серверов, один из которых вышел из строя, вы потеряли 5% информации — это неприятно, но, так как это случайная выборка, не все данные потеряны полностью. При этом, если наблюдалась какая-то закономерность, у вас могут быть проблемы. Например, если на сломавшийся сервер обычно поступала информация из конкретного географического региона, вы можете лишиться несоразмерного объема данных по этому отдельному региону, что может существенно повлиять на результаты анализа.

Возможны и другие сценарии, при которых выборка окажется необъективной. Например, представьте, что вы проводите опрос среди своих клиентов и даете респондентам две недели на то, чтобы прислать ответы. Ответы, полученные после указанной даты, рассматриваться не будут. А теперь предположим, что из-за проблем с доставкой группа клиентов получила свои заказы с опозданием. Возможно, они недовольны этой ситуацией и хотели бы выразить свое мнение, также ответив на ваш опрос и прислав его даже с опозданием. Если вы не учтете их ответы при анализе данных, то можете исключить из выборки большую долю недовольных клиентов. Оставшаяся выборка будет нерепрезентативной. В своих обучающих материалах по статистике Дэниел Минтц приводит пример формирования необъективной выборки: «Вопрос, нравится ли вам участвовать в опросах: да или нет?»[1] Как вы думаете, кто примет участие в этом опросе, а кто нет?

Причина, по которой пропущены данные, чрезвычайно важна. (Далее мы воспользуемся терминологией из области статистики, хотя она и ужасна.) Необходимо изучить, являются ли данные:

MCAR

Пропуски совершенно случайны, например распределяемый случайным образом трафик веб-сервера.

MAR

Пропуски случайны, но есть закономерности. Пропущенные данные — это функция от *наблюдаемых*, непропущенных данных, например веб-сервер, обслуживающий определенный регион, результатом чего стало уменьшение размера выборки почтовых индексов.

[1] URL: https://www.youtube.com/watch?v=zP638EdC0N4.

MNAR

Пропуски неслучайны, а пропущенные данные — функция других *пропущенных* данных, например недовольные покупатели и их ответы на опрос. Это наиболее опасный случай, где присутствует серьезная необъективность.

Чем ниже по списку, тем больше у вас может возникнуть сложностей и тем меньше шансов справиться с ситуацией.

Самое важное — понимать, что может послужить источником необъективности. В некоторых случаях можно намеренно ввести ограничения или проследить влияние на показатели. Как ни странно, бывают даже такие необычные ситуации, при которых пропущенные предвзятые данные могут не оказать никакого влияния на показатели.

Когда я преподавал статистику, то приводил следующий пример, чтобы показать свойства медианного значения. Есть такой необычный спорт — голубиная гонка. Владельцы почтовых голубей отвозят своих питомцев за сотни миль от дома, выпускают, а затем мчатся домой и ждут их возвращения. Так как это «гонка», то по возвращении каждого голубя фиксируется время, за которое он долетел до дома: например, голубь номер шесть вернулся через два часа три минуты, голубь номер одиннадцать — через два часа тринадцать минут и так далее. Неизбежно некоторые голуби не возвращаются: возможно, они сбились с курса или стали жертвой хищников. Мы не можем вычислить *среднее* время возвращения всех птиц, так как по некоторым из них нет данных. При этом, если больше половины вернулись, можно вычислить *медианное* значение времени полета. Нам известна величина выборки, известна продолжительность времени полета более половины участников выборки, мы знаем, что все пропущенные данные будут меньше значения последней прилетевшей птицы. Таким образом, мы вполне можем вывести медианное значение: оно будет достоверным с этим набором пропущенных данных. Иногда выбор правильных показателей может спасти ситуацию (выбору системы показателей посвящена глава 6).

ДУБЛИРОВАНИЕ ДАННЫХ

Еще одна распространенная проблема — дублирование данных. Это означает, что одна и та же запись появляется несколько раз. Причины могут быть разными: например, предположим, у вас десять файлов, которые нужно внести в базу данных, и вы случайно

загрузили файл номер шесть дважды, или при загрузке файла возникала ошибка, вы остановили процесс, устранили ошибку и повторили загрузку, но при этом первая половина данных загрузилась в вашу базу дважды. Дублирование данных может возникнуть при повторной регистрации. Например, пользователь прошел регистрацию несколько раз, указал тот же самый или другой адрес электронной почты, в результате чего у него появилась другая учетная запись с той же самой персональной информацией. (Звучит просто, но подобная неопределенность может оказаться весьма коварной.) Дублирование информации также может возникнуть в результате того, что несколько приборов фиксируют ее по одному событию. В исследовании медицинских ошибок, о котором шла речь ранее, в 35% случаев причиной ошибки был неправильный перенос данных из одной системы в другую: иногда данные терялись, иногда дублировались. По данным госпиталя Джонса Хопкинса, в 92% случаев дублирование информации в их базе данных происходило в момент регистрации стационарных больных.

Когда речь идет о базах данных, есть несколько способов предотвратить дублирование. Наиболее эффективный — добавление ограничений в таблицу с базой данных. Вы можете создать составной ключ, который определяет одно или несколько полей и делает запись уникальной. После добавления этого ограничения у вас будет появляться оповещение, если вводимая комбинация данных совпадет с уже существующей в таблице. Второй способ — выбор варианта загрузки данных по принципу «все или ничего». Если в момент загрузки данных обнаруживается проблема, происходит откат на изначальные позиции, а новая информация в базе данных не сохраняется. Это дает шанс разобраться с причиной проблемы и повторить процесс загрузки данных без дублирования информации. Наконец, третий (менее эффективный) подход — выполнять две операции при загрузке: первая операция — SELECT, чтобы выяснить, не присутствует ли уже такая запись, вторая операция — INSERT, добавление новой записи.

Подобное дублирование данных случается чаще, чем вы думаете. Если вы не знаете, что в ваших данных встречается продублированная информация, это может повлиять на ваши показатели. Но хуже всего, что в какой-то момент времени это все равно обнаружится. А если качество данных будет поставлено под сомнение хотя бы однажды, это снизит доверие к выводам аналитиков, и эти выводы не будут учитываться в процессе принятия бизнес-решений.

УСЕЧЕННЫЕ ДАННЫЕ

При загрузке информации в базу данных часть ее может потеряться (Anderson → anders или 5456757865 → 54567578). В лучшем случае можно лишиться пары символов в форме обратной связи. В худшем может произойти усечение и объединение идентификационных данных двух разных клиентов и вы непреднамеренно объедините данные двух разных клиентов или заказов в один.

Как такое может произойти? В обычных реляционных базах данных при создании таблицы задаются название и тип каждого поля: например, должен быть столбец под названием «Фамилия» с ячейками, содержащими до 32 символов, или столбец «ID клиента» с целым числом в диапазоне от 0 до 65535. Проблема в том, что не всегда заранее известно максимальное количество символов или максимальное значение идентификатора, с которыми вам придется столкнуться. Возможно, вы получите образец данных, рассчитаете длину ячейки и для подстраховки увеличите это значение в два раза. Но вы никогда не узнаете наверняка, достаточно ли этого, пока не начнете работать с реальными данными. Более того, в базах ошибки с усечением данных, как правило, относятся к категории *предупреждений*: появляется оповещение, но процесс загрузки данных не прекращается. В результате такие проблемы легко не заметить. Один из способов предотвратить это — изменить настройки в базе данных, чтобы предупреждения отображались как полноценные ошибки и заметить их было легче.

ЕДИНИЦЫ ИЗМЕРЕНИЯ

Еще один источник проблем с качеством данных — несовпадение единиц измерения, особенно когда речь идет о международных командах и наборах данных. CNN сообщает[1]:

> Агентство NASA потеряло орбитальный аппарат по исследованию Марса стоимостью 125 млн долл. из-за того, что команда технических специалистов корпорации Lockheed Martin использовала при расчетах английские единицы измерения [фунт-секунда], в то время как специалисты самого агентства пользовались более привычной метрической системой [ньютон-секунда] для управления аппаратом.

[1] URL: http://edition.cnn.com/TECH/space/9909/30/mars.metric.02/.

Да, это действительно настолько важно. Единственный способ избежать подобного — иметь четко налаженную систему коммуникации. Разработайте нормативный документ, утверждающий процедуру всех проводимых измерений, то, как они должны выполняться, и в каких единицах измерения должен указываться результат. Необходимо, чтобы документ был однозначным и не допускал иных толкований, а итоговая база данных сопровождалась подробным словарем базы данных.

Другая область, где единицы измерения имеют критическое значение, — денежные валюты. Представим сайт для электронной коммерции, на котором размещен заказ стоимостью 23,12. В США по умолчанию будет считаться, что это 23,12 долл., в то время как во Франции это будет 23,12 евро. Если заказы из разных стран окажутся объединены в одну базу данных учета информации по валютам, то итоговый анализ будет иметь отклонения в сторону более слабой валюты (поскольку в числовом выражении цена за тот же предмет будет выше) и фактически окажется бесполезен.

Базы данных должны обеспечивать столько метаданных и контекста, сколько необходимо, чтобы избежать подобного недопонимания.

Кроме того, можно просто принять метрическую систему и придерживаться ее (проснись, Америка!).

ЗНАЧЕНИЯ ПО УМОЛЧАНИЮ

Следующая проблема с данными, которую в некоторых случаях бывает сложно отследить, это значения по умолчанию (рис. 2.3A и D). Пропущенные данные могут отражаться в базе данных как NULL, но также может использоваться определенное значение, которое можно задать. Например, 1 января 1900 года — стандартная дата по умолчанию. С ней могут быть разные проблемы. Во-первых, если вы забудете о том, что эта дата появляется по умолчанию, результаты анализа могут вас весьма озадачить. Предположим, вы оставили это значение по умолчанию в ячейке с датой рождения. Аналитиков может смутить тот факт, что столько людей в вашей базе данных старше 100 лет. Во-вторых, при неудачном значении по умолчанию есть риск перестать различать пропущенные и актуальные данные. Например, если вы устанавливаете «0» как значение по умолчанию для пропущенных данных, а значение актуальных данных тоже может быть равным 0, впоследствии вы не сможете определить, в какой ячейке отражены результаты измерения, а в какой просто пропущены данные. Отнеситесь к выбору значений по умолчанию внимательно.

Происхождение данных

При обнаружении проблемы с качеством данных важно отследить источник данных. В этом случае можно будет извлечь из анализа проблемную выборку или предложить более эффективные процессы и протоколы работы с этими данными. Для метаданных, хранящих информацию об источнике данных и историю их изменений, я использую термин «происхождение данных».

Эти метаданные делятся на два типа: *история источников* (отслеживает, откуда появились данные) и *история преобразований* (отслеживает, какие изменения претерпевали данные).

В моей команде мы, например, ежедневно собираем файлы данных от разных разработчиков и загружаем их в нашу базу данных для проведения анализа и составления отчетов. Обычно промежуточные таблицы, в которые мы заносим всю информацию, содержат два дополнительных поля: время начала загрузки (конкретного файла или группы файлов) и название файла. Таким образом, если у нас возникают проблемы с качеством данных, мы легко можем определить, из какого файла эти данные, и уточнить их у разработчиков. Это пример *истории источников*.

В транзакционных базах данных (то есть тех, которые поддерживают работающие приложения и используются, например, для обработки заказов, а не для составления отчетов) довольно часто встречаются два поля: `created_at` (время создания) и `last_modified` (последнее изменение). Как следует из названия полей, они содержат уточняющую информацию о времени создания записи (эта метаинформация заносится один раз и больше не меняется) и о времени, когда было сделано самое недавнее изменение (эта метаинформация обновляется в режиме реального времени каждый раз, когда в запись вносятся любые изменения). Иногда в таблице может быть дополнительное поле `modified_by`, в котором фиксируется имя пользователя, внесшего последнее изменение. Это помогает определить, например, было ли изменение в заказе или адресе электронной почты сделано самими пользователями или представителем, действующим от имени клиента. В данном случае элемент `created_at` — история источников, в то время как элементы `last_modified` и `modified_by` отражают историю преобразований. Наиболее детальный инструмент отслеживания происхождения — таблицы с журналом событий, где четко протоколируется, какие именно изменения, кем и когда были внесены.

Метаданные о происхождении должны быть элементом проактивной стратегии проверки, поддержания и улучшения качества данных.

Велика вероятность, что важность фактора происхождения данных будет только расти. Сегодня становится все легче создавать системы для сбора и хранения собственных данных и предлагать для коммерческого использования подходящие дополнительные данные от третьих сторон (такие как демографические данные по почтовым индексам или история покупок по адресам электронной почты). Этим компаниям необходимо создавать более обширный контекст вокруг своих клиентов, а также вокруг своих открытых и внутренних данных по событиям и транзакциям. Это требует создания объектов на основе многочисленных источников данных, а также изменения существующих данных, например восстановления пропущенных данных или пояснения данных дополнительными характеристиками, такими как предполагаемый пол, цель и так далее. При этом всегда должна оставаться возможность отследить первоначальные значения данных, их источник, а также причину или метаинформацию по любому изменению данных.

Качество данных как совместная ответственность

Причины, обусловливающие снижение качества данных, могут быть самыми разными. Помимо уже перечисленных ранее, могут возникнуть проблемы с определением окончания строк, проблемы с кодировкой, когда данные в кодировке Юникод сохраняются в ASCII (это происходит сплошь и рядом), могут быть поврежденные данные, усеченные файлы, несовпадения в именах и адресах (см. табл. 2.1). Вопросами качества данных должны заниматься не только специалисты по сбору и обработке данных — эту ответственность должны разделять все сотрудники компании.

Разработчик внешнего интерфейса может добавить в форму на сайте функцию контроля правильности ввода почтового индекса. Специалист по обработке данных может добавить контрольную цифру при передаче данных в другое хранилище. Администратор базы данных может проверить и предотвратить дублирование информации или отследить ошибки при загрузке данных. Однако сложно ожидать, что им известно, какие показатели систолического артериального давления находятся в пределах нормы, а какие нет. Когда компания получает данные на основе заполненных форм, руководители подразделений, эксперты в предметных областях и аналитики должны быть в тесном

Таблица 2.1. Краткий обзор некоторых типов проблем с качеством данных и потенциальные варианты их решения. Более подробный список можно найти у Singh and Singh. A descriptive classification of causes of data quality problems in data warehousing, IJCSI Intl. J. Comp. Sci 7, no. 3 (2010): 41–50

Аспект	Проблема	Решение
Точность	*Ввод данных*: вставка символа	*Веб*: выпадающее меню, автозаполнение. *Аналог*: двойной ввод
Точность	*Ввод данных*: удаление символа	*Веб*: выпадающее меню, автозаполнение. *Аналог*: двойной ввод
Точность	*Ввод данных*: изменение символа	*Веб*: выпадающее меню, автозаполнение. *Аналог*: двойной ввод
Точность	*Ввод данных*: перестановка символов местами	*Веб*: выпадающее меню, автозаполнение. *Аналог*: двойной ввод
Точность	*Ввод данных*: недопустимые значения	*Веб*: проверка формы на соответствие. *База данных*: ограничение ячейки
Точность	*Ввод данных*: формат даты	*Веб*: автоматическая вставка формата даты. *База данных*: словарь базы данных, унификация (например, в формате ГГГГ-ММ-ДД)
Точность	Дублирующиеся записи	*База данных*: ограничения в виде комбинации клавиш, устранение повторов
Точность	Повреждение данных	Контрольная цифра или контрольная сумма
Точность	Разная кодировка данных (например, в одной таблице данные в кодировке UTF-8, а в другой — в ASCII) или ошибки при смене кодировки (например, в кодировке ASCII имя Jose может сохраниться как Jos)	*База данных*: стандартизация на базе одной широко принятой кодировки, например Latin1 или UTF-16
Точность/взаимосвязанность	Усечение значения	*База данных*: увеличение поля для ввода данных, смена статуса предупреждений на ошибки
Взаимосвязанность	Объединение ячеек (например, «Doe, Joe» может быть сложно вставить в другую таблицу, где этот же человек записан как «Joe Doe»)	*Приложение или база данных*: используйте отдельные ячейки

Аспект	Проблема	Решение
Взаимосвязанность	Разные первичные ключи для одной информационной единицы в разных системах осложняют правильное объединение данных	*Приложение или база данных*: унифицированная система идентификации
Непротиворечивость	Противоречивые данные (например, разные адреса одного человека в разных системах)	*База данных*: центральная пользовательская система, решение на основе установленного правила, какие данные более надежные
Путаница	Противоречивые временны́е зоны	*Веб*: автоматический выбор времени. *База данных*: словарь базы данных, унификация (например, на основе всемирного координированного времени (UTC)
Путаница	Заполнение ячеек другими данными (например, использование пустой ячейки `middle_name` для сохранения статуса заказа)	*Приложение или база данных*: наиболее эффективные методы, четкая, прописанная схема работы
Путаница	Путаница с кодировкой (например, HiLowRangeTZ3)	*База данных*: словарь базы данных
Путаница	Двусмысленные пропущенные данные (например, означает ли значение «0» пропущенные данные или актуальное значение «0»?)	*Приложение или база данных*: выбор разумных значений по умолчанию вне пределов возможных значений
Полнота	Частичные ошибки при загрузке	*База данных*: оповещения (возвращение к начальной стадии до загрузки)
Полнота	Пропуски совершенно случайны (MCAR)	*Анализ*: выборка с запасом, веса для категорий
Полнота	Пропуски случайны, но есть закономерности (MAR): данные пропущены как функция *наблюдаемых* или непропущенных данных	*Анализ*: ограничение анализа до того, когда данные можно использовать безопасно
Полнота	Пропуски неслучайны (MNAR): пропущенные данные — функция других *пропущенных* данных	*Анализ*: изменение или повторение процесса сбора данных

Аспект	Проблема	Решение
Полнота	Неправильное число или разделитель данных, вызывающие появление дополнительных столбцов или удаление	Поля данных (Quote fields), проверка качества источника данных
Своевременность	Устаревшие данные из-за медленных обновлений (например, журнал изменений адресов)	Более быстрая и качественная обработка данных
Происхождение	Сложность в определении, когда или почему было изменено значение	*Приложение или база данных*: более качественное отслеживание изменений, добавление в базу данных ячеек для фиксирования происхождения

контакте с разработчиками внешнего интерфейса, чтобы допустимые границы ввода данных были заданы правильно. Кроме того, они должны принимать участие в процессе формулирования требований и управления проектом, чтобы обеспечить контроль качества данных там, где это возможно. Как уже отмечалось ранее, специалисты по аналитике должны активно участвовать в процессе сбора данных.

Далее руководители направлений и эксперты в предметных областях должны проверить качество данных. Аналитики должны провести разведочный анализ или воспользоваться собственными методами определения, находятся ли значения в допустимых границах, соблюдаются ли ожидаемые закономерности (например, соотношение систолического и диастолического давления), оценить объем пропущенных данных и так далее. На фермерском рынке шеф-повар ресторана сам выбирает продукты, пробует авокадо, нюхает базилик. Образно говоря, это его сырые ингредиенты. У аналитиков должно быть такое же отношение к данным. Это их сырые ингредиенты, которые они должны тщательно отобрать.

Руководители направлений, как правило, принимают решения о покупке баз данных у третьих сторон, о разработке инструментов по сегментированию аудитории в ходе опроса клиентов или о проведении A/B-тестирования онлайн. Они тоже должны задумываться об объективности данных, на которые опираются. Они должны проводить сами или делегировать проведение разведочного анализа данных, составлять диаграммы распределения и обнаруживать «пятидюймовых» людей.

ГЛАВА 3

Сбор данных

https://t.me/it_boooks/2

Ошибки, возникающие при использовании неправильных данных, все же меньше, чем те, которые возникают при отсутствии данных.

Чарльз Бэббидж[1]

Сложно даже представить себе ту власть, которой может обладать человек, когда в его распоряжении столько информации самого разного рода.

Тим Бернерс-Ли[2]

В предыдущей главе мы обсудили вопросы качества данных и их правильного сбора. В этой главе фокус сместится на выбор правильных источников для сбора данных и предоставления специалистам по аналитике. Мы остановимся на следующих вопросах: как расставить приоритеты при выборе источников данных, как осуществить сбор данных, как определить ценность данных для компании.

Собирайте все что можно

Предположим, вы внедряете новый процесс оформления и оплаты заказов на сайте. Вас интересует, *как именно он работает* по сравнению с вашими показателями. Для этого вы можете проанализировать конверсию, размер корзины и другие параметры. Кроме того, вам было бы весьма полезно понять, как этот новый процесс *воспринимается со стороны покупателей*. Например, на некоторых сайтах добавление товара в корзину происходит в один клик мыши, так что модель поведения покупателя может быть следующей: он добавляет в корзину все, что его заинтересовало,

[1] Чарльз Бэббидж (1791–1871) — английский математик, изобретатель первой аналитической вычислительной машины. *Прим. перев.*

[2] Тим Бернерс-Ли (р. 1955) — британский ученый, создатель Всемирной паутины. Автор множества разработок в области информационных технологий. *Прим. перев.*

а перед оформлением заказа делает окончательный выбор, удаляя лишнее. На других сайтах добавление товаров в корзину и удаление из нее происходит не так просто, и фактически покупателю нужно принять окончательное решение перед добавлением товара в корзину. Очевидно, что всестороннее изучение и измерение процесса оформления и оплаты заказов помогает лучше его понять и внести изменения или улучшения.

В своей книге Building Data Science Teams[1] Ди Джей Патиль отмечает:

> Легко сделать вид, что вы действуете на основании анализа данных. Но если на самом деле собирать и измерять все доступные вам данные и думать о том, что означают собранные вами данные, вы намного опередите все те компании, которые лишь заявляют об управлении на основе данных.

Собирайте все доступные данные. Никогда не знаешь, какая информация может понадобиться, а шанс собрать данные часто выдается только один, и вы будете кусать локти, когда поймете, что нужная вам информация больше недоступна. Чем больше данных вы соберете, тем больше вероятность, что вам удастся смоделировать и понять поведение пользователей (как в примере с процессом оформления и оплаты заказа) и, что более важно, понять *контекст* их действий. Контекст — наше все. Таким образом, чем лучше компания поймет своих покупателей, их вкусы, намерения, желания, тем успешнее ей удастся улучшить пользовательский опыт своих клиентов благодаря персонализации, рекомендациям или совершенствованию сервиса, что будет способствовать возникновению так называемого длинного хвоста[2].

При разработке онлайновых продуктов сбор абсолютно всех данных нельзя считать чем-то уникальным. Вы контролируете источник данных: сбор информации относительно одной какой-то характеристики может проводиться с помощью того же самого или похожего механизма, что и сбор информации относительно другой характеристики. То есть существует возможность использования общих шаблонов, потоков данных и механизмов хранения. Компания, в которой действительно уделяется большое внимание данным, вероятно, будет характеризоваться более широким горизонтом мышления. В такой компании все остальные функции также окажутся организованы на основе дан-

---

[1] Подробнее о книге: http://www.oreilly.com/data/free/building-data-science-teams.csp.

[2] Anderson C. The Long Tail: Why the Future of Business Is Selling Less of More. New York: Hachette Books, 2005. Издана на русском языке: Андерсон К. Длинный хвост. Эффективная модель бизнеса в Интернете. М. : Манн, Иванов и Фербер, 2012. *Прим. ред.*

ных: маркетинг, продажи, обслуживание клиентов, цепочка поставок, работа с персоналом. Если по каждому из этих направлений имеется набор внутренних и внешних *источников данных* в разных форматах, с разным временем ожидания, проблемами с качеством данных, с разными требованиями к безопасности и соответствия нормативам и так далее, то это начинает превышать возможности команды специалистов по работе с данными. Это тот случай, когда «собирать все что можно» звучит как отличная идея, которая оборачивается серьезной «головной болью», когда доходит до дела.

Более того, этот процесс требует финансовых затрат. Чем больше данных, тем лучше[1] (см. приложение А, где приведены примеры и объяснение, почему это так), но какую цену компания за это платит? На создание инфраструктуры для сбора, очистки, трансформации и хранения данных нужны средства. Компания несет издержки на поддержание работоспособности этой инфраструктуры, резервное копирование данных, интеграцию источников этих данных для обеспечения целостной картины бизнеса. Кроме того, возможны значительные дальнейшие издержки на обеспечение качественного инструментария для специалистов по анализу данных, чтобы они могли максимально эффективно использовать эти несопоставимые источники данных. Компании не обойтись без всего этого, если она стремится, чтобы правильные данные попали в руки специалистов по анализу.

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ БОЛЬШИХ ДАННЫХ

Специалисты по большим данным выделяют три аспекта сбора и обработки большого количества данных: объем, разнообразие и скорость[2].

Объем

Объем данных напрямую влияет на издержки на их хранение и изменения. Хотя абсолютно верно, что расходы на хранение данных снижаются экспоненциально[3] (сегодня хранение информации обходится в 0,03 долл. за GB по сравнению с примерно 10 долл. за GB

[1] Fortuny E. J. de, Martens D. and Provost F. Predictive Modeling with Big Data: Is Bigger Really Better? Big Data 1, no. 4 (2013): 215–226. URL: http://online.liebertpub.com/doi/full/10.1089/big.2013.0037.

[2] Впервые встречается у Д. Лейни. 3D Data Management: Controlling Data Volume, Velocity and Variety. Application Delivery Strategies by META Group Inc., February 6, 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[3] URL: http://www.mkomo.com/cost-per-gigabyte-update.

в 2000 году), число доступных источников данных повысилось настолько значительно, что это перекрывает снижение затрат на хранение информации.

Разнообразие

Это еще один важный аспект данных. С одной стороны, разнообразный набор источников способен обеспечить более богатый контекст и более полную картину. Таким образом, прогноз погоды, данные по инфляции, сообщения в социальных медиа могут оказаться весьма полезными для понимания продаж ваших продуктов. При этом, чем разнообразнее тип данных и источники данных (CSV-файлы из одного источника, объекты JavaScript (JSON) из другого источника, почасовой прогноз погоды отображается здесь, а данные о запасах — здесь), тем выше будут издержки на интеграцию. Довольно сложно собрать все данные вместе, чтобы получить общую картину.

Скорость

Объем данных, который требуется обработать в единицу времени. Представьте, что в ходе дебатов кандидатов в президенты вам нужно проанализировать сообщения в Twitter, чтобы вывести общее настроение избирателей. Необходимо не только обработать огромный объем информации, но также оперативно предоставить обобщенную информацию о настроении нации относительно комментариев во время дебатов. Масштабная обработка данных в режиме реального времени — процесс сложный и дорогостоящий.

(В некоторых случаях компании выделяют еще один аспект — «достоверность», для характеристики качества данных.)

Даже компаниям, сегодня собирающим огромные объемы данных, например Facebook, Google и Агентству национальной безопасности США (NSA), на это потребовалось время. Только со временем удается выстроить источники данных, взаимосвязи между ними и возможности обработки данных. Требуется рациональная и тщательно продуманная *стратегия* обеспечения данными. Более того, в большинстве компаний команды, работающие с данными, ограничены в ресурсах: они не в состоянии делать все и сразу, так что им приходится расставлять приоритеты, с какими источниками данных работать в первую очередь. Реальность такова, что процесс сбора данных идет медленно и последовательно: всегда возникают непредвиденные задержки

и проблемы, так что приходится сосредоточиваться на ценности, рентабельности инвестиций и влиянии, которое новый источник данных окажет на компанию. Этому и будет посвящена данная глава.

Расстановка приоритетов при выборе источников данных

В обычных малых или средних компаниях, ограниченных в ресурсах, специалистам по работе с данными, как правило, приходится выбирать, с каким источником данных работать. Чем они при этом руководствуются? Определяя приоритеты при выборе источников данных, компания, в которой управление осуществляется на основе данных, должна сосредоточиться на таком важном аспекте, как *ценность* данных для бизнеса.

Основная цель команды по работе с данными заключается в том, чтобы предоставлять данные, отвечающие потребностям определенных подразделений компании и их аналитиков, и помогать оказывать влияние на эффективность деятельности компании. У каждой команды или подразделения, как правило, имеется набор «основных» данных. Например, для специалистов по обслуживанию клиентов это могут быть данные по взаимодействию с ними посредством электронной почты, телефонных звонков, социальных медиа, данные по заказам клиентов, а также разбор конкретных ситуаций. На основе этих данных команда может выполнять свои основные функции — максимально эффективно обслуживать клиентов. Кроме того, специалисты могут объединить эти источники для создания целостного взгляда на сценарии взаимодействия с клиентами. Они могут предоставить обобщенные показатели продуктивности работы команды, такие как среднее время решения проблемы клиента, а также проанализировать тип взаимодействий в случае каждого источника. У каждой команды специалистов должны быть свои основные данные. Однако, помимо этого, у них могут быть и другие данные, способные дополнить основной набор. Например, коэффициент дефектности продукции или данные A/B-тестирования, проясняющие, какая новая характеристика товара привела клиентов в замешательство. На основе этих данных специалисты могут прогнозировать частоту и характер ситуаций при работе с клиентами, которых можно ожидать. Эти другие источники данных также могут быть ценными и оказывать влияние, но они не критические.

Проблема компании с ограниченными ресурсами в том, что команда специалистов по работе с клиентами — лишь одна из многих.

У команд специалистов в других областях есть свои наборы основных данных и свои пожелания относительно информации, «которую было бы неплохо иметь». Специалист по работе с данными или руководитель команды по работе с данными вынужден уравновешивать все эти запросы от разных команд специалистов. В табл. 3.1 приводится ряд показателей, способных помочь в расстановке приоритетов. Основной фактор — рентабельность инвестиций (ROI), но стоит принимать во внимание и другие факторы, такие как доступность, полнота, качество данных и некоторые другие.

Таблица 3.1. Аспекты, на которые следует обратить внимание при расстановке приоритетов при выборе новых источников данных в условиях ограниченности ресурсов

Приоритетность	**Причина**	**Объяснение**
Высокая	Данные нужны незамедлительно	Если у какого-то подразделения компании есть острая потребность в данных и жестко установленный срок, данные для этого подразделения нужно подготовить в первую очередь и как можно быстрее
Высокая	Данные обеспечат высокую ценность	Например, если данные могут помочь значительно увеличить прибыль или снизить издержки, обеспечив таким образом высокую ROI, то этот источник данных должен иметь высокий приоритет
Высокая	Разным командам требуются одни и те же данные	ROI повышается, если вы одновременно в состоянии удовлетворить запросы нескольких направлений бизнеса
Высокая	Данные имеют краткосрочный или потоковый характер	Некоторые интерфейсы API потоковых социальных медиа или аппаратных устройств обеспечивают лишь кратковременное окно для получения данных, после которого данные утрачиваются навсегда
Средняя	Дополнение существующего набора данных для повышения его качества	Новые данные дополняют существующий набор данных и обеспечивают значительно более богатый контекст (подробнее это будет обсуждаться далее)

Приоритетность	Причина	Объяснение
Средняя	Специалисты по работе с данными могут повторно использовать код обработки существующих данных	Если команда знакома с источником или его интерфейсом прикладного программирования (API) и способна использовать существующий код, это снижает вероятность неизвестных переменных или неожиданностей
Средняя	Данные легко доступны	Иногда запрос на получение данных может возникнуть просто потому, что удобный клиент Python или API делают процесс сбора данных очень простым, или данные обладают четкой и простой структурой. Если с этим источником данных можно разобраться относительно быстро и он обладает некоторой очевидной ценностью, возможно, стоит воспользоваться им по-быстрому
Средняя	Удобный интерфейс прикладного программирования (API) позволяет собрать данные за прошлые периоды	Если данные не требуются срочно и вы точно знаете, что всегда сможете получить к ним доступ, тогда, вероятно, стоит обратить внимание на более важные источники данных. Например, если вам потребуются необработанные данные Google Analytics для создания архива, вы всегда сможете их получить
Низкая	У аналитиков есть какой-то доступ к данным и обходные пути для их получения	Если у аналитиков есть хоть какой-то доступ к данным, пусть не идеальный, например через дашборд, и есть возможность выгрузить эти данные через CSV или другими способами, тогда приоритетность этого источника низкая. Вероятно, есть другие источники данных, к которым у компании пока нет доступа, но которые могут иметь для компании бóльшую ценность
Низкая	Низкое качество данных	Если в качестве данных есть сомнения, то их использование в лучшем случае ничего не даст, а в худшем будет контрпродуктивным

Приоритетность	Причина	Объяснение
Низкая	Данные необходимо извлекать из веб-страниц	Так как владельцы сайтов часто изменяют HTML и CSS веб-страниц и они не всегда хорошо структурированы, подобная обработка данных может оказаться довольно сложной и потребовать усилий
Низкая	Низкая вероятность того, что данные будут использоваться	Если это данные из категории тех, которые «хорошо было бы иметь», и для них нет четкого применения, это не самый хороший выбор

Очевидно, что самые разные, нередко конкурирующие аспекты определяют, какой новый источник данных целесообразно использовать в компании. Существует тонкий баланс между издержками на приобретение новых данных и сложностью этого процесса и той ценностью, которую эти данные имеют для аналитиков и компании в целом.

Установление взаимосвязи

Очевидно, что для проведения более глубокого анализа важное значение имеет сбор данных внутри компании: вы получаете определенные данные из отдела маркетинга, данные из отдела продаж, данные по цепочке поставок. Однако еще бóльшую ценность эти данные обретают, когда вы начинаете устанавливать взаимосвязи между смежными данными. Что я имею в виду?

Представьте, что вам предложили тысячу элементов для составления пазла, но на коробке при этом нет изображения того, что должно в итоге получиться. По мере сортировки элементов вы выделили группу элементов голубого цвета. Вероятно, это небо. Группа элементов зеленого цвета может изображать траву. Вот вы нашли глаз. Но чей — животного или человека? У вас появляется смутное представление о картинке в целом, но не хватает деталей. Детали возникают, когда вы начинаете соединять смежные элементы, например элементы с изображением глаза и элементы с изображением уха. Появилась ясность. Давайте рассмотрим эту ситуацию с точки зрения аналитики.

Предположим, вы пользуетесь сервисом Google Analytics для анализа того, как пользователи попадают на ваш сайт. Вы получаете подборку

веб-страниц, с которых произошел переход на ваш сайт, а также список поисковых запросов, географию пользователей и так далее, что дает вам общее представление о выборке пользователей или генеральной совокупности (это условные «кусочки неба»). Вы анализируете результаты опроса покупателей за последние три месяца: 75% респондентов нравится цена, 20% похвалили качественное обслуживание и так далее (это «кусочки травы»). У вас складывается общее представление о состоянии дел, но весьма поверхностное, так как данные остаются разрозненными.

Теперь, наоборот, представим, что мы имеем дело с одним заказом (см. рис. 3.1). Белинда Смит заказывает комплект садовой мебели. Если сопоставить ее заказ с сессией, во время которой она совершила покупку, можно сделать определенные выводы: она потратила 30 минут на просмотр 15 разных комплектов садовой мебели, прежде чем остановилась на одном. Очевидно, у нее не было четкого представления, какой комплект она ищет. Как она попала на страницу компании? Если добавить сопутствующую информацию, выяснится, что она ввела поисковый запрос в Google и перешла на сайт компании. Это подтверждает наше предположение относительно ее пользовательского поведения. Если к этому добавить полную историю ее онлайновых покупок, можно сделать вывод, что Белинда часто покупает товары для дома, а за последний месяц количество таких покупок у нее резко увеличилось. Те факты, что Белинда часто совершает покупки онлайн и пользуется поисковым сервисом Google, позволяют предположить, что у нее нет лояльности к конкретным брендам и компании придется постараться, чтобы она совершила повторную покупку. Каждый раз, добавляя новый элемент информации на *индивидуальном* уровне, вы начинаете лучше понимать этого покупателя. Продолжим. На основе данных переписи населения США определим вероятный пол по имени: Белинда практически наверняка женщина. Отлично. При оплате покупки она указала адрес доставки. Попробуем извлечь демографические данные на основании индекса. Это пригород с большими земельными участками, где живут состоятельные люди. Как еще можно проверить этот адрес? «Пробьем» его по единой базе данных недвижимости (MLS). Интересно, база данных показывает, что это дом с бассейном. Эту информацию можно использовать для полезных рекомендаций. Что еще? Дом был продан всего шесть недель назад. Ага, вероятно, Белинда только что въехала в новый дом. По результатам другого проведенного нами анализа известно, что новоселы часто покупают коврики, кровати и лампы (да, так и есть, я сам проводил

этот анализ). Наконец, она нажала на виджет «приведи друга», чтобы получить купон при оформлении заказа. Так как она приняла условия пользовательского соглашения с Facebook, это открыло ее социальную сеть. (Подробнее о вопросах этики и сохранения конфиденциальности мы поговорим в главе 12.)

Для аналитика этот подробный профиль и контекст предлагают огромный объем сырых данных, с которыми можно работать. Специалист получает четкое представление о демографических данных клиента, истории его покупок и, в этом случае, даже о его мотивации. Проведите такой анализ для других ваших клиентов и автоматизируйте хотя бы часть этого анализа — и вы получите значительное стратегическое преимущество.

Установление взаимосвязи между элементами информации на этом индивидуальном уровне, в противоположность уровню сегмента, имеет огромную ценность и должно влиять на решения о том, какой набор данных использовать следующим (без нарушения этических норм и границ конфиденциальности), а также как связать эти данные с уже имеющимися на индивидуальном уровне.

Сбор данных

Теперь, когда мы разобрались, какие данные нужно собирать, давайте кратко остановимся на вопросе, как это делать.

В случае со многими источниками можно просто *системно* собирать все доступные данные. Есть много способов управления потоками данных. Можно воспользоваться интерфейсом прикладного программирования (API) или собирать файлы с FTP-сервера, можно даже проводить анализ экранных данных и сохранять что необходимо. Если это одноразовая задача, с ней легко справиться. Однако при частом обновлении или добавлении данных нужно решить, как работать с этим потоком. Для небольших таблиц или файлов может быть проще полностью заменять их новым, более масштабным набором данных. В моей команде маленькими у нас считаются таблицы с количеством строк до 100 тысяч включительно. Для работы с более крупными массивами данных необходимо установить более сложный процесс с анализом изменений. В самом простом случае новые данные всегда вносятся в новые ряды (например, журналы транзакций, где не должно быть обновлений или удалений текущих данных). В этом случае можно просто добавить (INSERT) новые данные в таблицу с текущими данными.

В более сложных случаях необходимо решить, будете ли вы добавлять (INSERT) строку с новыми данными, удалять (DELETE) или обновлять (UPDATE).

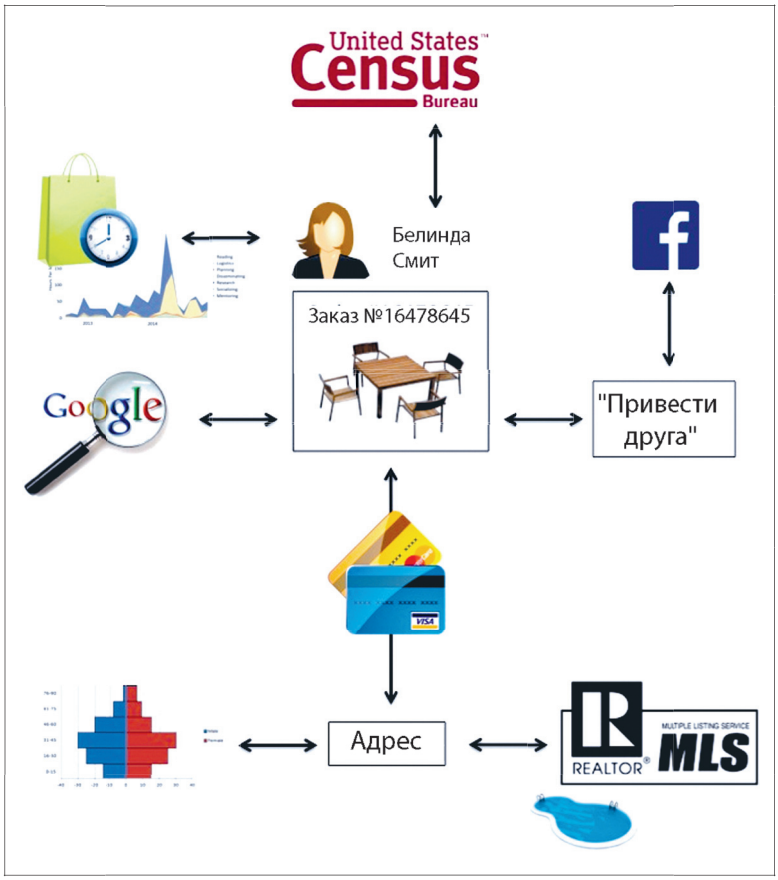


Рис. 3.1. Определение более широкого контекста для заказа Белинды на основе разных источников данных

Источник: https://www.slideshare.net/CarlAnderson4/ddo-seattle

Для других источников данных может потребоваться сделать выборку. Проведение опросов и обработка результатов иногда бывает слишком дорогостоящим процессом, так же как и проведение клинических исследований или анализ всех записей в Twitter. То, каким образом осуществляется выборка, оказывает огромное влияние на качество данных. Мы поговорим об этом подробнее в главе 8, однако необъективная выборка в значительной степени влияет на качество данных

и возможность их использования. Самый простой подход заключается в формировании «простой случайной выборки»[1], когда данные, которые будут включены в выборку, определяются простым подбрасыванием монетки. Суть в том, чтобы выборка была действительно репрезентативной относительно более крупного массива данных, из которого она формируется.

Внимательно стоит отнестись к формированию выборки данных, которые собираются в течение определенного периода времени. Предположим, вам требуется выборка сессий сайта за день. Вы отбираете 10% сессий и загружаете информацию о них в базу данных для последующего анализа. Если вы проделываете эту процедуру ежедневно, у вас формируется набор независимых сессий, выбранных случайным образом, но при этом вы можете упустить данные о пользователях, которые посетят сайт в последующие дни. То есть в выборке может не оказаться информации о пользователях с несколькими сессиями: они могут попасть в выборку в понедельник, но не попадут туда при их возвращении на сайт в среду. Таким образом, если вас больше интересуют последующие повторные сессии, а пользователи вашего сайта часто возвращаются, для вас может быть эффективнее выбрать случайным образом посетителей и отслеживать их сессии на протяжении определенного времени, чем делать случайную выборку сессий. В этом случае вы получите для работы данные более высокого качества. (Хотя, возможно, вам будет не слишком приятно наблюдать за пользователями, которые не возвращаются на сайт.) Механизм формирования выборки должен определяться тем бизнес-вопросом, ответ на который вы ищете.

И последнее: следует ли собирать сырые или агрегированные данные? Некоторые поставщики данных предлагают дашборды, где данные агрегированы в соответствии с ключевыми показателями, необходимыми аналитикам. Для аналитиков это может оказаться большим подспорьем. Однако если данные действительно ценные, для аналитиков такого подхода будет недостаточно: они непременно захотят еще больше углубиться в их изучение и рассмотреть их с самых разных сторон, а с дашбордами сделать это не удастся. Все эти отчеты и дашборды эффективно использовать для архивного хранения данных. В других случаях, как показывает мой опыт, лучше по возможности собирать сырые данные, так как вы всегда сможете агрегировать их согласно показателям, но не наоборот. Имея сырые данные, вы сможете

[1] Подробную информацию можно найти по ссылке: https://en.wikipedia.org/wiki/Simple_random_sample.

работать с ними как вам потребуется. Конечно, бывают редкие случаи, когда сбор сырых данных нерационален, например в силу большого их объема и высокой стоимости хранения или по причине того, что поставщик данных предлагает ценный сервис для обработки этих показателей (что вы не сможете сделать самостоятельно), но в большинстве случаев сбор сырых данных все-таки предпочтителен.

Покупка данных

Как правило, внутренние системы сбора данных в компании обеспечивают огромные массивы информации, которые можно дополнить данными, находящимися в открытом доступе, хотя иногда нужно заплатить за получение дополнительных данных от третьих сторон.

Существует множество причин, по которым вам может потребоваться покупать данные. Ранее мы анализировали заказ Белинды Смит на комплект садовой мебели, чтобы показать значимость контекста. Во-первых, другие партнеры, поставщики или даже государственные структуры могут располагать данными, способными обеспечить нужный контекст и добавить в вашу головоломку смежные элементы. Во-вторых, вы можете обладать внутренними данными, но данные третьей стороны могут выигрывать по объему или качеству.

В некоторых случаях выбор мест, где приобретать данные, может оказаться ограниченным. Например, единая база данных недвижимости (MLS) практически монопольно предоставляет информацию по сделкам. В других случаях возможна прямая конкуренция. Например, данные по профилям клиентов на основании их покупок, оплаченных с помощью кредитных карт, можно приобрести у нескольких компаний: Datalogix, Axciom, Epsilon или Experian. Это рыночные условия в действии.

При выборе между несколькими источниками данных, например при приобретении базы данных, в которой почтовые индексы соотнесены с местностью на карте, необходимо принять во внимание несколько факторов, в том числе перечисленные ниже.

Цена

Аналитики и их боссы любят «халяву», но иногда стоит заплатить за данные высокого качества. Следует взвесить, насколько рациональна цена и какой ценностью эти данные обладают для компании. Подробнее об этом мы поговорим в следующем разделе.

Качество

Насколько чисты и надежны эти данные?

Эксклюзивность

Подготовлен ли этот набор данных исключительно для вас и получите ли вы с его помощью преимущество перед конкурентами?

Выборка

Можно ли получить выборку, которая позволит судить о качестве и характере данных, а также понять формат без необходимости предварительно брать на себя обязательства?

Обновления

Насколько часто данные меняются или устаревают? Насколько часто данные обновляются?

Надежность

При обращении к интерфейсу прикладного программирования (API) каково время работоспособности системы? Каковы ограничения по обращениям к API или по другим сервисным соглашениям?

Безопасность

В случае, если данные важны, осуществляется ли их шифровка и какие меры безопасности предпринимаются при передаче?

Условия использования

Есть ли условия лицензирования или другие ограничения, которые могут не позволить воспользоваться данными в полной мере?

Формат

У всех есть любимые форматы данных, тем не менее обычно предпочтительно использование форматов, удобных для восприятия человеком, таких как CSV, JSON или XML (это подразумевает исключение бинарных форматов, кроме стандартного сжатия), так как эти форматы более удобны для использования при проведении анализа. Наконец, насколько просто вам будет поддерживать этот формат? Не потребуется ли от вас дополнительных вложений и времени на работу с этим форматом?

Документация

Предпочтение следует отдавать источникам, способным предоставить документацию. Обычно стоит поинтересоваться, как осуществляется сбор данных (чтобы понять, насколько они надежны и представляют ли они ценность для компании) и есть ли словарь данных (в нем указываются поля, тип данных, примеры значений и другая важная бизнес-логика, включенная в значения этих полей; см. табл. 3.2). Рэндалл Гроссмен, CDO корпорации Fulton Financial, заметил: «Словарь данных, которому можно доверять, — это самое важное, что CDO может предложить бизнес-пользователям».

Объем

Сможете ли вы обеспечить хранение большого объема данных? При этом ценные наборы данных не обязательно бывают большими. Например, почтовый индекс для расчетной рыночной территории (то есть территории охвата конкретного региона телевещанием, по оценке компании Nielsen Company) может иметь всего 41 тыс. строк, но эти данные могут быть очень полезны команде специалистов по маркетингу, оценивающей расходы на телевизионную рекламу.

Степень детализации

Подходят ли данные для анализа того уровня, который вам необходим?

Таблица 3.2. Пример словаря данных из проекта в области здравоохранения в Калифорнии

Название переменной в SAS (eHARS)	Ярлык	Описание	Значения	Формат SAS	Название переменной в HARS
aids_age_mos	Возраст при диагностировании СПИДа (ВИЧ, стадия 3), месяцы	Возраст при диагностировании СПИДа (ВИЧ, стадия 3) месяцы			age_mos
aids_age_yrs	Возраст при диагностировании СПИДа, годы	Возраст при диагностировании СПИДа (ВИЧ, стадия 3), годы			age_yrs

Название переменной в SAS (eHARS)	Ярлык	Описание	Значения	Формат SAS	Название переменной в HARS
aids_categ	Выявление случаев заболевания СПИДом	Выявление случаев заболевания СПИДом Центром по контролю и профилактике заболеваемости США (CDC) (ВИЧ, стадия 3), подсчитано на основе лабораторной информации и оппортунистических болезней у респондентов Для описания алгоритма, применявшегося для расчета aids_categ, см. часть 8 технического руководства eHARS	**7** — СПИД (ВИЧ, стадия 3), случай заболевания подтвержден на основе иммунологических критериев (число CD4-клеток или процент) **A** — СПИД (ВИЧ, стадия 3), случай заболевания подтвержден на основе критериев клинических болезней (ОИ) **9** — не СПИД (ВИЧ, стадия 3), случай заболевания не подтвержден	A_CAT, долл.	categ
aids_cdc	Выявление случаев заболевания СПИДом Центром по контролю и профилактике заболеваемости США (CDC)	Был ли случай заболевания СПИДом (ВИЧ, стадия 3) выявлен Центром по контролю и профилактике заболеваемости США? В этом случае он должен быть подтвержден на основе иммунологических критериев или критериев клинических болезней (aids_categ = A или 7)	**Y** — Да **N** — Нет	YN, долл.	N/A

Благодаря качественному словарю становится понятно, как определяются данные, в каком формате и с какими допустимыми значениями. В данном случае также очевидно, как эти данные используются программным обеспечением. Приведены несколько строк из eHARS[1] (Enhanced HIV/AIDS Reporting System — Улучшенная система сбора информации о ВИЧ/СПИДе) в Калифорнии. (SAS — статистический набор приложений, активно применяющийся в области медицины.)

Сколько стоит набор данных?

Посчитать, во сколько вам обходятся данные, относительно легко. Можно проанализировать величину прямых расходов на хранение (например, стоимость услуг Amazon Web Services), стоимость сервисов резервного копирования, зарплаты сотрудников, обеспечивающих хранение и управление данными, а также их непроизводственные расходы, плюс стоимость приобретения данных (если актуально). При этом компания с управлением на основе данных должна определить ценность этих данных для бизнеса. Какова их ROI? А вот это уже не так просто.

Д'Алессандро и др.[2] предложили фреймворк, позволяющий оценить прямую рентабельность инвестиций ROI в долларах, по крайней мере в определенных ситуациях. Они работают в сфере рекламы и разработали прогнозные модели для вычисления, какие рекламные объявления эффективнее всего показывать каждому пользователю. Они получают деньги только за переход пользователя по рекламному объявлению. При этом сценарии результат и выручка очевидны: они получают, скажем, 1 долл., если пользователь переходит по рекламному объявлению, и 0 долл., если пользователь ничего не делает. У них есть собственный набор данных, на основании которых они строят свои модели. Некоторые из них — ретроспективные, взятые на основе действовавших ранее цен, а некоторые были ими приобретены в прошлом (их относят к категории невозвратных затрат). Вопрос, которым они руководствуются: «Какова рентабельность моделей, построенных на наших собственных данных, по сравнению с моделями, построенными на данных от третьих лиц?» Для этого требуется определить три компонента:

---

[1] URL: https://github.com/d3/d3/wiki/Gallery.

[2] d'Alessandro B., Perlich C. and Raeder T. Bigger is Better, But At What Cost? Big Data 2, no. 2 (2014): 87–96. URL: http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0010.

1) какова стоимость действия (в данном случае действие — это переход пользователя, его стоимость — 1 долл.);
2) какова ожидаемая стоимость модели на основе наших собственных данных;
3) какова ожидаемая стоимость модели на основе наших данных и дополнительных данных третьей стороны.

Итого:
Стоимость данных = ожидаемая стоимость (модель на основе данных третьей стороны) – ожидаемая стоимость (модель без использования данных третьей стороны)

и
Предельная норма прибыли = стоимость (переход) × стоимость данных.

Предположим, у модели на основе собственных данных всего 1% вероятности, что по рекламному объявлению будет переход, а у модели на основе дополнительных данных третьей стороны эта вероятность составляет 5%. Ценность данных выше на 4%, а прирост ценности этих данных составляет 1 долл. × (5% – 1%) = 0,04 долл.

Располагая конкретным значением вроде этого, можно объективно определить целесообразность приобретения этих данных. Если стоимость дополнительных данных 0,04 долл., тогда это нерентабельно. А если их стоимость составит, например, 0,01 долл., решение очевидно.

Вы можете не ограничиваться только оценкой прироста ценности данных третьей стороны в дополнение к собственным данным. Когда речь идет о данных, в большинстве случаев самая важная роль отводится контексту. Д'Алессандро и др. провели интересный эксперимент, в ходе которого сравнили прирост ценности данных третьей стороны по сравнению со случайным таргетированием пользователей, то есть полным отсутствием данных по сравнению с данными только третьей стороны. Они получили положительный прирост ценности по целому ряду сегментов: стоимость по сегменту / 1 тыс. пользователей составила 1,8 долл. Затем они повторили эксперимент и использовали собственные данные плюс данные третьей стороны. Как вы думаете, какой результат они получили? Прирост ценности упал! Стоимость по сегменту на 1 тыс. пользователей теперь была около 0,02 долл. В контексте данных, которыми они уже располагали, дополнительные данные обеспечили положительную, но незначительно малую ценность (рис. 3.2), вероятнее всего, из-за избыточности данных.

Этот общий подход достаточно эффективен, так как есть возможность приобрести выборку данных, которую можно протестировать. Если

полученный результат хороший, можно приобрести полный набор данных. То есть они не связаны обязательством по приобретению полного набора данных, пока не проведут эксперименты, подтверждающие их ценность. К сожалению, не все поставщики данных и не всегда идут на такие условия. Тем не менее, возможно, вы вносите ежемесячную оплату за пользование данными. В таком случае вы можете проанализировать ценность данных с помощью описанных выше экспериментов и увидеть, насколько рентабельно их использование. Если для вас это нерентабельно, откажитесь от услуг этого поставщика.

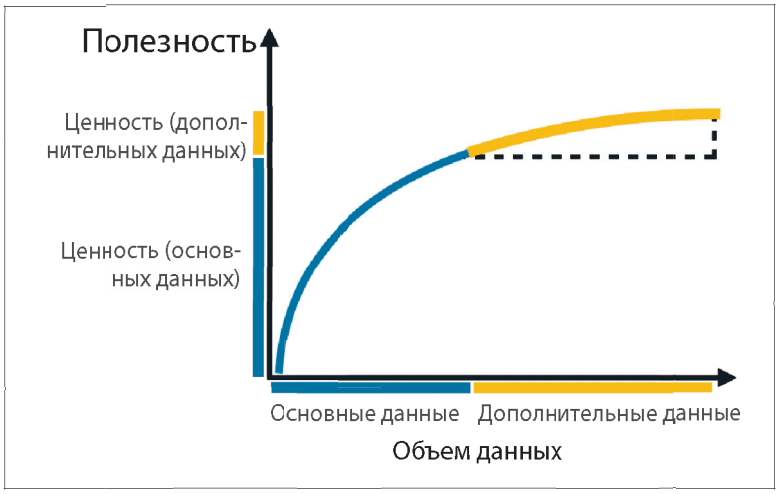


Рис. 3.2. Дополнительные данные должны способствовать повышению ценности, но наблюдается убывающая доходность

Источник: https://conferences.oreilly.com/strata/stratany2014/public/schedule/detail/37642

Авторы делают заключение:

> По мере того как большие данные превращаются в панацею при принятии многих решений по оптимизации бизнеса, для руководителей все большее значение приобретает способность рационально оценить свои решения и инвестиции в приобретение и использование данных. Без инструментов для проведения подобной оценки большие данные становятся скорее интуитивным подходом, чем научной практикой.

Аминь!

Хранение данных

Эта глава была посвящена нахождению и интеграции дополнительных данных. В результате этого процесса увеличивается объем данных, с которыми работают аналитики. При этом данные могут устаревать. Ранее мы уже говорили о стоимости данных — издержках на их приобретение, хранение и управление ими. Кроме того, есть издержки и риски, которые не так легко оценить: какой урон может нанести вашему бизнесу, например, утечка данных? Один из аспектов, о которых следует задуматься, — когда удалять данные (сокращая риск утечки и издержки на хранение) и когда перемещать данные на подходящий носитель для хранения.

У данных есть одна особенность: они множатся. Вы можете загрузить набор данных в реляционную базу, но на этом все не закончится. Ваши данные могут сохраниться в одну или несколько подчиненных баз при неполадках с сервером, на котором хранится основная база данных. И вот у вас уже две копии. Кроме того, вы можете проводить резервное копирование на сервер. Обычно таких резервных копий, на случай, если что-то пойдет не так, у вас может быть за несколько дней, даже за неделю. Так что вы теперь обладатель девяти копий, и хранение каждой из них стоит денег. Как поступить в такой ситуации? Один из вариантов — сопоставлять наборы данных с адекватным периодом ожидания, в течение которого их можно использовать или сохранить.

Рассмотрим такой пример: Amazon S3 — дешевый и простой способ хранения данных[1]. Хранение данных с помощью такого сервиса определенно обойдется дешевле, чем покупка и обслуживание дополнительного сервера для хранения резервных копий. Получить данные вы можете в любой момент, когда они вам потребуются. При этом Amazon также предлагает похожий сервис под названием glacier[2]. По сути, он очень похож на S3, но создавался как сервис для архивного хранения данных, и на получение данных может уйти четыре-пять часов. При текущем уровне цен стоимость glacier в три раза ниже, чем S3. В случае экстренной ситуации потребуются ли вам данные немедленно или вы сможете обойтись без них полдня или день?

Компании с управлением на основе данных следует тщательно оценить их стоимость. Изначально сосредоточиться нужно на основных данных, где любой простой может иметь серьезные последствия.

[1] URL: https://aws.amazon.com/ru/s3/.

[2] URL: https://aws.amazon.com/ru/glacier/.

Компании следует наладить процесс удаления устаревших данных (это бывает легче сказать, чем сделать) или, в крайнем случае, хотя бы перемещать эти данные на самые дешевые из возможных источников хранения.

Более эффективные компании с управлением на основе данных, например достигшие уровня прогнозного моделирования, могут разрабатывать модели, которые используют только самые необходимые данные и отбрасывают все остальные. Например, по словам Майкла Ховарда, CEO компании C9, «отдел продаж не хранит детали заказа более 90 дней»[1]. Если это так, то необходимо тщательно отбирать данные. Как мы показали, компании с управлением на основе данных следует стратегически подходить к выбору источников данных и к ресурсам компании на работу с данными. Аналитики выполняют важные функции по анализу потенциальных источников информации и поставщиков данных, по приобретению выборок и, по возможности, по оценке качества данных и применению выборки для определения ценности данных.

В следующей главе мы поговорим о самих специалистах по аналитике, об их функциях и о том, как можно организовать аналитическую работу в компании.

[1] URL: https://techcrunch.com/2014/09/06/three-marks-of-real-data-science/.

ГЛАВА 4

Специалисты по аналитике

По-настоящему хороший аналитик должен будоражить людей… Я знаю, что я первый получаю данные, а значит, я первый узнаю историю. Открывать что-то новое увлекательно.

Дэн Мюррей

Человеческий фактор — важный компонент компании с управлением на основе данных. Кто такие специалисты по аналитике и как должна быть организована их работа?

Эта глава посвящена специалистам по аналитике: разным их типам и навыкам, которыми они должны обладать. Мы рассмотрим самые разные позиции и познакомимся с людьми, которые их занимают. Кроме того, мы обсудим плюсы и минусы разных организационных структур для выполнения аналитической работы.

Типы специалистов по аналитике

В компании с управлением на основе данных, вероятнее всего, есть разные специалисты по аналитике, собранные в многочисленные команды. Есть разные описания этих аналитических позиций, и многие из перечисляемых навыков пересекаются. Я предлагаю собственную версию общего описания аналитиков, специалистов по работе с данными, бизнес-аналитиков, специалистов по обработке данных, по статистике, по количественному и экономическому анализу, финансовых аналитиков и специалистов по визуализации данных. Для каждого из этих типов специалистов я опишу навыки, которыми они должны обладать, инструменты, которыми они пользуются, а также приведу конкретные примеры. В вашей компании могут быть другие названия для этих специалистов, но без описанных навыков обычно невозможно эффективно работать с данными.

АНАЛИТИК

Это самый широкий и общепринятый термин, по крайней мере по сравнению с более узкими профессиональными ролями, о которых пойдет речь далее. В большинстве случаев их опыт можно условно представить в виде буквы «Т»: они обладают скромным опытом по целому спектру навыков, но очень глубокими знаниями и навыками в своей основной профессиональной области. В зависимости от своего профессионального опыта специалисты по аналитике могут быть как новичками, которые занимаются в основном сбором и подготовкой данных, так и высококвалифицированными аналитиками со специализацией по определенной теме. Такие аналитики часто бывают главными экспертами в разных областях, таких как работа с мнением клиентов, программы лояльности, электронный маркетинг, геоспециализированная военная разведка или отдельные сегменты фондового рынка. Конкретная роль в компании зависит от ее размера, зрелости, области специализации и рынка. В любом случае результат работы аналитика, скорее всего, будет представлять собой сочетание анализа и отчетов. Аналитики могут отличаться по степени владения техническими навыками и знания профессиональной области.

С одной стороны, есть аналитики, работающие исключительно в Excel и с помощью дашбордов. А с другой стороны, есть такие, как Самарт, который сам пишет программные коды на языке Scala для обработки большого объема сырых данных в компании Etsy. Изначально Самарт занимался политологией, а навыки аналитической работы получил в предвыборном штабе Барака Обамы во время работы в кампании 2012 года. Затем с помощью стандартной триады инструментов, наиболее популярных у аналитиков (R, SQL и Python), он начал проводить исследования в сети и с электронными рассылками. Сегодня он работает аналитиком в компании Etsy в Нью-Йорке, где продолжает проводить свои исследования, а также осуществляет анализ истории посещений пользователей и трендов, составляет отчеты и аналитические доклады. В компании он взаимодействует с продакт-менеджерами, техническими специалистами и дизайнерами и помогает им разрабатывать эксперименты, анализировать их с помощью Scala/Scalding, R и SQL и интерпретировать полученные результаты. Кроме того, он готовит общие аналитические отчеты для компании, а также более узконаправленные справки для руководителей, чтобы помочь им разобраться в трендах, поведении пользователей или других специфических вопросах.

Саманта — аналитик совсем другого рода. У нее степень бакалавра по бухгалтерскому учету, и она работает специалистом по данным в страховой компании Progressive Insurance в Кливленде, штат Огайо, в команде финансовых специалистов отдела по работе с исковыми заявлениями. Она занимается вопросами выморочного имущества (это категория наследуемого имущества, которая отходит государству в случае отказа от его получения), проводит аудит, анализ и проверяет соответствие законам штата в данной области. В ее работу входит подготовка отчетов и отслеживание собственности, от которой отказались, поиск интересных проектов, суммирование финансовых рисков, связанных с этими вопросами. В своей работе она использует такие инструменты, как SAS, Excel и Oracle, а также специализированные инструменты, такие как ClaimStation. От результатов ее работы зависит целый ряд аспектов, которыми занимаются другие специалисты в компании, в том числе это налог на прибыль корпораций, финансовые операции, ИТ, исковые заявления крупного бизнеса, а также исковые заявления отдельных людей. По словам Саманты, ее мотивирует, когда она «видит, что ее анализ приносит финансовую выгоду как компании, так и застрахованным у нас клиентам». В ее работе особенно важно внимание к деталям, поскольку она работает в жестко регулируемой отрасли, а в сферу ее обязанностей входит проверка деятельности компании на соответствие законам штата.

ИНЖЕНЕРЫ В ОБЛАСТИ ОБРАБОТКИ ДАННЫХ И АНАЛИЗА

Эти специалисты в первую очередь несут ответственность за сбор и обработку данных и перевод их в формат, удобный для проведения анализа. Они отвечают за аспекты операционной деятельности, такие как скорость обработки информации, масштабирование, пиковые нагрузки и ведение журнала операций. Кроме того, они могут отвечать за разработку инструментов, которые используют аналитики.

Знакомьтесь, это Анна. Во время подготовки диссертации по физике она поняла, что на самом деле ей интересно заниматься данными. Она окончила обучение с дипломом магистра и начала работать в компании Bitly в качестве специалиста по обработке данных. Анна занимается визуализацией больших объемов данных, обрабатывает данные с помощью набора инструментов Hadoop, внедряет алгоритмы машинного обучения. Затем она присоединилась к проекту Rent The Runway и сейчас работает там инженером по обработке данных. При помощи таких инструментов, как SQL, Python, Vertica, она поддерживает

инфраструктуру данных, на которой держится аналитический процесс, разрабатывает новые инструменты для повышения надежности данных, их своевременности и масштабируемости, а также взаимодействует с другими техническими специалистами компании, чтобы понимать любые изменения, которые они совершают и которые могут повлиять на данные.

БИЗНЕС-АНАЛИТИКИ

Эти специалисты обычно выступают связующим звеном между руководством (например, руководителями отделов) и технологическим отделом (например, разработчиками программного обеспечения). Их функции заключаются в улучшении бизнес-процессов или помощи в разработке новых или совершенствовании существующих бэкэнд- и фронтэнд-систем, например, в их функции входит улучшение воронки продаж на сайте.

Линн — старший бизнес-аналитик крупного интернет-магазина Macys.com. У нее степень бакалавра в области изобразительных искусств, опыт разработчика приложений, сертификат Профессионала в управлении проектами, кроме того, почти десятилетний опыт работы в области управления проектами и бизнес-аналитике, преимущественно в сфере книжной электронной коммерции. В функции Линн входит проведение анализа требований проекта, понимание потребностей клиентов, совершенствование бизнес-процессов, а также управление проектами, часто на основе гибкого подхода (Agile). Линн делится своими впечатлениями: «Ни один мой рабочий день не похож на другой. Сегодня я могу беседовать с пользователями на тему их ожиданий (то есть с предпринимателями, которые пользуются информационной системой управления товарами Масу), завтра я делаю обзор ответов пользователей вместе с разработчиками или отвечаю на вопросы разработчиков относительно ответов пользователей».

DATA SCIENTISTS (СПЕЦИАЛИСТЫ ПО РАБОТЕ С БОЛЬШИМИ ДАННЫМИ)

Этот широкий термин применяется для обозначения специалистов в области работы с большими данными, обладающих математическими или статистическими знаниями, обычно с более высоким уровнем образования в точных науках, а также развитыми навыками программирования. Мне нравится лаконичное определение Джоша

Уиллса: «Это человек, который разбирается в статистике лучше любого программиста и способен написать программный код лучше любого статистика»[1]. Тем не менее это не полное описание его функций, которые могут включать разработку «продуктов на основе данных», таких как рекомендательный сервис с применением машинного обучения, или прогнозное моделирование, или обработка естественного языка[2].

Трей — старший специалист по теории и методам анализа данных интернет-компании Zulily, расположенной в Сиэтле. Особенность этого интернет-магазина — ежедневные распродажи. У Трея степень магистра по социологии. Свое рабочее время Трей делит между самыми разными проектами — от разработки статистических моделей и рекомендательных алгоритмов для улучшения опыта пользователей до помощи менеджерам продуктов в интерпретации результатов A/B-тестирования. В основном он пользуется языком программирования Python (с такими библиотеками, как Pandas, Scikit-learn и Statsmodels), а также анализирует данные, используя SQL и системы управления базами данных Hive. Он обладает нужными техническими навыками для построения статистических моделей и считает способность доступно объяснить эти модели неспециалистам одним из важнейших качеств профессионала, занимающегося работой с данными. Любовь к обучению нашла отражение в его хобби: он ведет блог, в котором объясняет концепции работы с данными на примере данных по американскому футболу, а также рассказывает о том, как лучше понимать спортивную статистику[3].

СПЕЦИАЛИСТЫ ПО СТАТИСТИКЕ

Это квалифицированные сотрудники, которые занимаются в компании статистическим моделированием. Обычно у них не ниже степени магистра в области статистики, чаще всего они востребованы в таких сферах, как страхование, здравоохранение, исследования и разработки, государственное управление. Четверть всех специалистов по статистике в США работают на федеральное правительство, правительства штатов или органы местного самоуправления[4]. Часто они занимаются

---

[1] URL: https://twitter.com/josh_wills/status/198093512149958656.

[2] Conway D. The Data Science Venn Diagram, September 30, 2010. URL: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.
Anderson C. What is a data scientist? December 3, 2012. URL: http://www.p-value.info/2012/12/what-is-data-scientist.html.

[3] URL: http://thespread.us/.

[4] URL: https://www.bls.gov/ooh/math/statisticians.htm.

не только анализом данных, но и разработкой опросов, исследований, а также сбором протоколов для получения сырых данных.

Шон — специалист по статистике, поддерживающий проведение количественных маркетинговых исследований в офисе Google в Боулдере. У него степень бакалавра в области математики и научных вычислений и Ph.D.[1] в области статистики. Сегодня Шон также обеспечивает поддержку сотрудникам в других командах, часто при возникновении необходимости переходя из проекта в проект. С одной стороны, он может заниматься сбором, очисткой, визуализацией и оценкой качества данных из нового источника. А с другой стороны, он опирается на свои технические навыки для разработки алгоритмов кластеризации, чтобы улучшить онлайновые геоэксперименты по поиску, разработать байесовские модели временны́х рядов или оценить уровень индивидуального просмотра на основе данных домохозяйств с помощью алгоритма Random Forests. В основном он пользуется средой R, особенно для анализа и визуализации данных (в частности, такими пакетами, как ggplot2, plyr/dplyr и data.table). Помимо этого он применяет в своей работе языки программирования типа SQL и пользуется Python и Go.

КВАНТЫ

Специалисты по количественному анализу, как правило, обладают хорошей математической подготовкой и обычно работают в финансовом секторе, моделируя управление риском и движение фондового рынка со стороны как покупателей, так и продавцов. Например, пенсионный фонд может нанять кванта, чтобы тот сформировал оптимальный портфель облигаций, способный покрыть будущие обязательства фонда. Квантами могут стать бывшие математики, физики или технические специалисты. Некоторые из них — особенно аналитики алгоритмической торговли (самые высокооплачиваемые специалисты из всех аналитиков) — обладают уверенными навыками программирования на таких языках, как C++, они способны обрабатывать данные и предпринимать действия с крайне небольшим временем ожидания.

Сатиш — квант в компании Bloomberg в Нью-Йорке. У него глубокие знания в области прикладной математики и проектирования

[1] Ph.D. (лат. Philosophiae Doctor, доктор философии) — ученая степень, которая присуждается в западной системе высшего образования. Эта степень не имеет никакого отношения к философии (кроме исторического) и присуждается во всех научных областях. По разным мнениям, эта степень соответствует степеням кандидата или доктора наук в нашей стране (или находится между ними). *Прим. ред.*

электрических систем, о чем свидетельствует его степень Ph.D. Он пользуется средой R (ggplot2, dplyr, reshape2), языком программирования Python (scikit-learn, pandas) и Excel (для сводных таблиц) для построения самых разных статистических моделей, а затем при помощи C/C++ запускает некоторые из них. Эти модели часто определяют относительную ценность различных категорий активов с фиксированной доходностью. Помимо этого, он выступает в роли внутреннего консультанта, и ему приходится решать самые разные задачи — от кредитных моделей для ценных бумаг с ипотечным покрытием до прогнозирования объема ветровой энергетики в Великобритании. По его словам, «огромный объем финансовых и аналитических данных, доступный для специалистов Bloomberg, беспрецедентен для отрасли. Поэтому нас воодушевляет осознание того, что большинство предлагаемых нами моделей имеют ценность для всех наших клиентов». Одна из сложностей работы с финансовыми данными заключается в том, что у них очень «длинный хвост», и таким образом в моделях необходимо тщательно учитывать эти редкие, нестандартные события.

СПЕЦИАЛИСТЫ ПО ЭКОНОМИЧЕСКОМУ АНАЛИЗУ И ФИНАНСОВЫЕ АНАЛИТИКИ

Специалисты, которые занимаются внутренней финансовой отчетностью, аудиторскими проверками, прогнозированием, анализом эффективности производственной деятельности и так далее. У Патрика степень бакалавра по философии, политологии и экономике, а также опыт работы в качестве специалиста по анализу рынков заемного капитала в компании RBS Securities. Сейчас он занимает позицию менеджера по розничному финансированию и стратегии в компании Warby Parker в Нью-Йорке, где отвечает за планирование и анализ финансов в розничной сети, а также разработку стратегии по открытию новых магазинов. Он проводит много времени, работая с Excel, управляя прибылями и убытками склада и ключевыми показателями результативности (KPIs), разрабатывая модели будущей деятельности, изучая отклонения в моделях и проводя анализ развития рынка. Сегодня Патрик тратит около 60% рабочего времени на подготовку отчетов, а оставшееся время — на проведение анализа, тем не менее это соотношение увеличивается в пользу времени на аналитическую работу по мере того, как улучшается его знакомство с инструментами бизнес-аналитики в компании и повышаются навыки работы с этими инструментами.

СПЕЦИАЛИСТЫ ПО ВИЗУАЛИЗАЦИИ ДАННЫХ

Это люди с развитым чувством прекрасного, которые создают инфографику, дашборды и другие графические элементы. Кроме того, они могут заниматься написанием программного кода при помощи JavaScript, CoffeeScript, CSS и HTML и работают с библиотеками визуализации данных, такими как D3 (эффективная и красивая библиотека визуализации, описанная в книге Скотта Мюррея Interactive Data Visualization for the Web) и HTML5.

Джим (Джим В., см. рис. 4.1) получил степень магистра в области теории и практики вычислительных систем со специализацией в сфере биоинформатики и машинного обучения. Он работал в компании Garmin, где создавал графические пользовательские интерфейсы для навигационных устройств. После этого в биологическом научно-исследовательском институте он проводил анализ масштабной последовательности данных. Именно тогда он познакомился с библиотекой визуализации данных D3 и начал вести блог, посвященный этой теме, где публикует доступные и понятные руководства для пользователей. Сегодня Джим занимает пост специалиста по визуализации данных и специалиста по теории и методам анализа данных в лаборатории данных корпорации Nordstrom в Сиэтле. В своей работе он использует такие инструменты, как Ruby, Python и среду R (в частности пакеты ggplot2 и dplyr). Он обеспечивает поддержку систем персонализации и рекомендаций, а также осуществляет визуализацию данных. Основными его «клиентами» становятся сотрудники из других подразделений компании. В крупных компаниях иногда могут быть дополнительные специалисты, которые занимаются исключительно подготовкой отчетов или применением определенного инструмента бизнес-аналитики. Другие специалисты могут работать только с инструментами обработки и анализа больших данных, например Hadoop или Spark.

Как вы сами видите, названия специалистов, работающих с данными, как и их функции, во многом пересекаются. В основном они обрабатывают данные с помощью разных языков программирования типа SQL.

В одних случаях требуются более серьезные навыки программирования, а в других можно обойтись и без них. Нередко требуется построение статистических моделей с применением SAS или R. В большинстве случаев работа аналитика объединяет подготовку отчетов и собственно проведение анализа.

Аналитика — это командный спорт

Аналитика требует слаженной командной работы. В компании с управлением на основе данных, в которой четко налажены рабочие процессы, присутствуют как аналитики разных типов, так и сотрудники с дополняющими их навыками. При найме новых сотрудников принимается во внимание «портфолио» совокупных навыков всей команды, чтобы найти таких потенциальных кандидатов, которые «закроют» и усилят проблемные области.

Например, на рис. 4.1 приведен профиль команды лаборатории по работе с данными компании Nordstrom в 2013 году. Легко можно определить сильнейших математиков и статистиков в команде (Элисса, Марк и Эрин), сильнейших разработчиков (Дэвид и Джейсон В.), а также специалиста по визуализации данных (Джим В., о котором шла речь ранее). Я поинтересовался у директора лаборатории Джейсона Гоуэнса, что он думает насчет расширения команды, на что он ответил: «Во-первых, мы придерживаемся «правила двух пицц» Джеффа Безоса[1], а потому количество членов нашей команды вряд ли сильно изменится. Мы уверены, что такой подход помогает нам сконцентрироваться на том, что нам кажется серьезными возможностями. Во-вторых, каждый член команды привносит в нее что-то уникальное, что помогает расти всем остальным».

Еще в момент формирования команды они поступили весьма мудро, наняв сильного специалиста по визуализации данных, хотя многие другие команды делают этот шаг гораздо позже. Наличие красиво оформленных и подтвержденных концепций, основанных на данных, помогло команде лаборатории утвердить свой авторитет в рамках всей компании. «Джим очень помог нам вызвать интерес к нашей работе у остальных сотрудников, с помощью своих навыков визуализации данных он буквально вдохнул жизнь в то, что мы делаем», — говорит Джейсон.

Как уже отмечалось, профессиональные знания и навыки специалистов по теории и методам анализа данных, которые часто приходят в коммерческий сектор из академической среды, условно можно изобразить в виде буквы «Т». А если у эксперта две основные области специализации — то в виде числа пи (π). Найм новых сотрудников и формирование команд можно назвать «аналитическим тетрисом».

[1] Джефф Безос — основатель и генеральный директор Amazon. Его «правило двух пицц» гласит: группа должна быть настолько малочисленной, чтобы ее можно было накормить всего двумя пиццами. Обычно это команда из пяти-семи человек. *Прим. перев.*

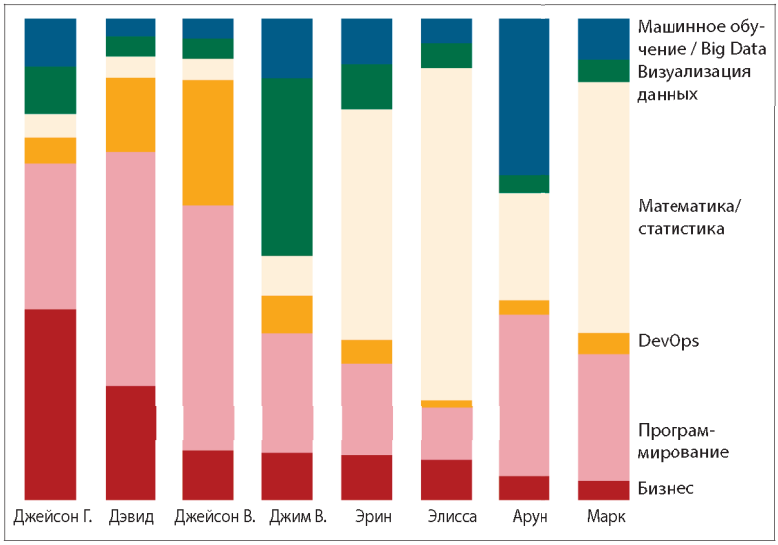


Рис. 4.1. Профиль команды лаборатории данных компании Nordstrom (по состоянию на 2013 год). МО = машинное обучение. DevOps — относительно новый термин, обозначающий интеграцию разработки и эксплуатации программного обеспечения

В 2012 году Харрис и др.[1] провели опрос среди нескольких сотен специалистов по работе с данными и разделили их на пять групп по ключевому навыку, как они сами себя охарактеризовали:

— бизнес;

— математика / анализ операций;

— машинное обучение / большие данные;

— программирование;

— статистика.

Они выделили четыре кластера ролей.

Предприниматели

Специалисты по работе с данными, у которых лучше всего развиты навыки, связанные с ведением бизнеса (форма буквы «Т»), и в меньшей степени развиты остальные навыки.

[1] Этому посвящена книга Analyzing the Analyzers. URL: http://www.oreilly.com/data/free/analyzing-the-analyzers.csp.

Исследователи

Специалисты, у которых лучше всего развиты навыки по работе со статистикой и в меньшей степени — навыки в области машинного обучения / больших данных, бизнеса и программирования.

Разработчики

Эксперты с двумя областями специализации (форма числа Пи) — с сильными навыками в сфере программирования и машинного обучения / больших данных и умеренными навыками по трем оставшимся категориям.

Творческие специалисты

Специалисты, «которые в среднем не считаются ни самыми сильными, ни самыми слабыми ни в одной из групп по ключевому навыку».

Профили этих четырех ролей представлены на рис. 4.2. Легко отметить широкое разнообразие среди этих четырех типов.

Таблица 4.1. Соответствие аналитических ролей, перечисленных ранее в этой главе, и ролей, выделенных Харрисом и др. (2013)

Предприниматель	**Творческий специалист**	**Разработчик**	**Исследователь**
Бизнес-аналитик Аналитик Специалисты по экономическому анализу и финансовые аналитики	Специалист по визуализации данных	Data Scientist Инженеры в области обработки данных и анализа	Специалист по статистике Квант

Эти четыре роли примерно соответствуют названиям позиций специалистов по работе с данными (табл. 4.1). В более крупных и сложно организованных компаниях можно выделить больше ролей, в компаниях малого бизнеса, вероятно, меньшее количество специалистов будет выполнять более широкие функции. Кроме того, стоит отметить, что, хотя Харрис и др. назвали творческих специалистов «ни самыми сильными, ни самыми слабыми ни в одной из групп по ключевому навыку», они не выделили при этом визуализацию и коммуникацию в отдельную категорию по ключевому навыку, хотя это чрезвычайно

важные навыки для команды. Проблема с данными также заключается в слабости опросов: они ограничены теми категориями, которые изначально предлагают авторы исследования. В данном случае было важно понять, что творческие специалисты — часть успешных команд, но нет ясности относительно их вклада в общий успех.

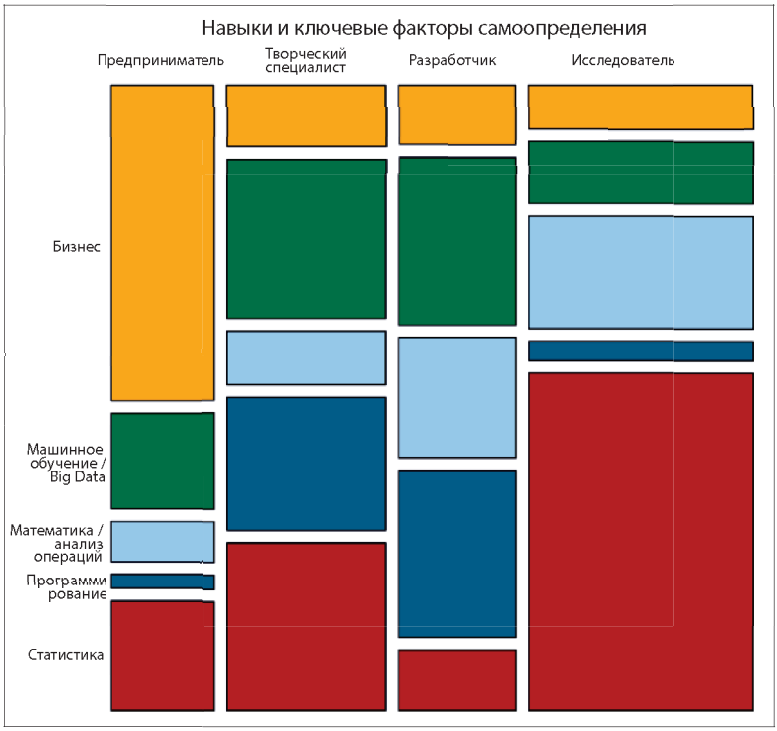


Рис. 4.2. Профиль навыков четырех кластеров респондентов

Источник: Харрис и др., 2013, рис. 3.3

В идеале при найме новых сотрудников руководителю следует принять во внимание три уровня.

Индивидуальный

Насколько подходит кандидат? Обладает ли он нужными навыками, потенциалом и стремлением, которые ищет компания?

Командный

Насколько кандидат впишется в команду и сможет ли закрыть слабые места?

Рабочий

Насколько профиль команды соответствует поставленным перед ней задачам? То есть каким должен быть профиль команды, чтобы она оптимально выполняла поставленные перед ней задачи? Например, если задача главным образом состоит в разработке финансовых прогнозных моделей, то состав команды будет отличаться от того, который требуется, если задача заключается в оптимизации процесса обслуживания клиентов.

Навыки и качества

Какие качества определяют хорошего аналитика?[1]

Аналитический склад ума

Он не обязательно должен иметь научную степень по математике или статистике, но его не должна пугать, по крайней мере, описательная статистика (медиана, мода, квартиль и так далее, см. главу 5), и он должен быть готов обучаться.

Внимание к деталям и методичность

Если эти цифры, отчеты и результаты анализа попадают на стол к руководителю и влияют на принятие бизнес-решений, лучше, если они будут правильными. И лучше, если аналитик всегда будет придерживаться правила «семь раз отмерь, один отрежь».

Рациональный скептицизм

Хороший аналитик интуитивно понимает, когда что-то не так с сырыми или агрегированными данными или результатами анализа. Во-первых, он прогнозирует, какие значения были бы более вероятны. Во-вторых, ставит под сомнение качество данных, еще раз проверяет их источник и расчеты, когда показатели отклоняются от ожидаемых.

Уверенность в себе

Аналитик презентует результаты своей работы коллегам (руководителям). Если эти результаты неожиданные или отражают

[1] Подробное обсуждение этого вопроса можно найти в книге Стивена Фью Now You See It (Analytics Press), с. 19–24.

неэффективность в каких-то аспектах деятельности, коллеги могут поставить их под вопрос, а потому аналитик должен обладать уверенностью в себе, чтобы отстаивать свою точку зрения.

Любопытство

Частично задача аналитика состоит в том, чтобы извлекать из информации полезные для бизнеса уроки и выводы, так что он постоянно должен проявлять любопытство, выдвигая разные гипотезы и тестируя интересные аспекты данных.

Навыки общения и повествования

Работа аналитика теряет всякий смысл, если ее результаты не передаются людям, принимающим решения, которые способны ими воспользоваться. Аналитику необходимо уметь рассказать увлекательную и связную историю на основе данных и результатов анализа. Для этого он должен обладать навыками визуализации данных и уметь убедительно формулировать свои мысли в устной и письменной форме (подробнее об этом в главе 7).

Терпение

Многие факторы находятся вне зоны контроля аналитика, в том числе точность или доступность источника данных, утерянные данные, меняющиеся требования, скрытая необъективность в данных, которая становится очевидной только после выполнения анализа и приводит к необходимости переделывать все заново. Без терпения здесь не обойтись.

Любовь к данным

Точно так же, как многим программистам просто нравится процесс написания кода, некоторым людям информация нравится как ресурс, благодаря которому им удается понять окружающий их мир и оказать на него влияние. Им просто нравится пытаться во всем разобраться досконально. Нанимайте таких людей.

Стремление учиться

Это качество присуще не только аналитикам. Успеха добиваются те, кто стремится узнавать новое, следит за новостями в своей профессиональной области, учится, чтобы совершенствовать свои знания и навыки.

Прагматизм и деловой подход

> Аналитик должен уметь концентрироваться на правильных вопросах. Иногда бывает трудно удержаться, чтобы не свалиться в «кроличью нору» и не потратить кучу времени на изучение отдельного пограничного случая, который не окажет никакого влияния на бизнес. Подобно хорошему редактору, аналитик всегда должен держать в голове общую картину и точно знать, в какой момент нужно остановиться и переключиться на что-то другое, чтобы более эффективно потратить свое время.

Я спросил у Дэниела Танкеланга, отвечающего за качество поиска в социальной сети LinkedIn, чем он руководствуется при найме на работу аналитиков. Он ответил:

> По моему мнению, аналитику необходимы три качества. Во-первых, он должен быть умным, способным неординарно решать задачи и не только обладать аналитическими навыками, но и знать, как и когда их применять. Во-вторых, он должен быть не просто теоретиком, а демонстрировать, что у него есть и способность, и горячее желание реализовывать свои решения на практике посредством подходящих инструментов. В-третьих, у него должно быть понимание того продукта, с которым он работает, основанное на опыте или интуиции, он должен уверенно ориентироваться в этой области и ее проблемах, и он должен задавать правильные вопросы.

Кен Рудин, глава аналитики социальной сети Facebook, уверен[1]:

> С помощью науки, технологий и статистики можно найти ответы, но по-прежнему большим искусством остается умение задавать правильные вопросы… Сегодня недостаточно нанимать людей с научной степенью в области статистики. Нужно быть уверенным, что у этих людей есть деловая хватка. Мне кажется, деловой подход становится самым важным активом и критическим навыком, которым должен обладать каждый аналитик.
>
> Как понять, есть ли у кандидата на позицию аналитика это качество? В ходе собеседования не концентрируйтесь только на том, как рассчитать тот или иной показатель.

[1] URL: https://www.youtube.com/watch?v=RJFwsZwTBgg.

Предложите потенциальному сотруднику практический случай из вашего бизнеса и спросите, на какие показатели он бы обратил внимание в этом конкретном случае. Вам все будет ясно из его ответа.

Еще один инструмент

С точки зрения практических навыков, без всяких сомнений, большинство аналитиков во всем мире использует в своей работе Microsoft Word, Excel и PowerPoint в качестве основных инструментов. Они доказали свою эффективность. Тем не менее поразительно, как может сказаться на продуктивности применение нескольких дополнительных инструментов.

Далее мы рекомендуем вам бросить вызов. Если вы аналитик, бросьте вывоз самому себе: в течение следующего месяца или квартала освойте еще один инструмент или программу. Если вы руководите аналитиками, поставьте перед ними такую задачу. Попробуйте и увидите, какой будет результат. Вы будете удивлены.

Стоит обратить внимание на следующие аспекты.

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ И СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

R представляет собой популярную среду для осуществления статистических вычислений и располагает исключительными библиотеками визуализации данных (такими как ggplot2)[1]. Например, можно прочитать данные в формате CSV и визуализировать отношения между всеми возможными парами переменных с помощью всего двух команд:

```
данные<-read.csv(имя_файла.csv);
pairs(данные)
```

На рис. 4.3 показан результат действия этих двух команд. Во второй панели верхней строки отражена взаимосвязь между шириной чашелистика (ось *x*) и длиной чашелистика (ось *y*) цветков ириса.

[1] URL: https://www.r-project.org/.

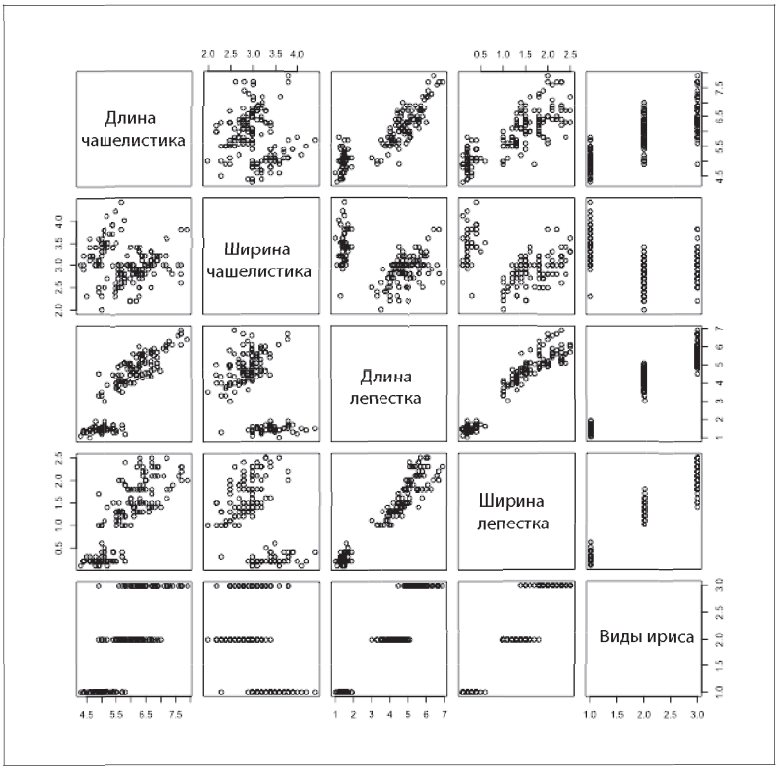


Рис. 4.3. Результат применения команд (относительно задачи по ирисам) в среде R. Речь идет о наборе данных относительно 150 экземпляров ириса, по 50 экземпляров из трех видов, который собрал ботаник Эдгар Андерсон и сделал знаменитым Рональд Фишер[1]. Корреляция между переменными и разница между тремя видами становится очевидной, если рассмотреть все взаимоотношения в совокупности, как на рисунке

Таким образом, этот инструмент может стать чрезвычайно полезным для быстрого проведения разведочного анализа данных. (Не менее популярны и эффективны неоткрытые SAS и SPSS.) Всего около 6700 пакетов для любых типов данных, моделей, областей и визуализации. Это открытые источники, доступные бесплатно[2]. Если вы уже знакомы со средой R, то можете освоить новый пакет R и расширить свои навыки.

[1] URL: https://en.wikipedia.org/wiki/Iris_flower_data_set.

[2] Об эффективных инструментах с открытым исходным кодом можно узнать из книги П. Джанерта Data Analysis with Open Source Tools (O'Reilly).

ЗАПРОСЫ К БАЗАМ ДАННЫХ

В то время как Excel может быть очень эффективным инструментом, при работе с ним иногда возникают проблемы, связанные с обработкой большого объема данных: при определенном объеме данных и применении функции ВПР (VLOOKUP) программа может сильно затормозить работу компьютера. Именно поэтому язык программирования SQL — ценный инструмент в наборе любого аналитика. Этот язык можно назвать относительно стандартизированным, несмотря на незначительные отличия в языке в разных базах данных (таких как MySQL, PostgreSQL и Access). Так что если вы знакомы с ним, это обеспечит вам свободу переключения между разными реляционными базами данных. Вы сможете делать запросы к базам данных независимо от объема данных (обрабатывать миллионы строк), делиться запросами с коллегами (делиться небольшими текстовыми запросами, а не огромными массивами сырых данных). Кроме того, вы сможете обеспечить воспроизводимость процесса (можно легко повторить процесс анализа еще раз).

Есть множество книг, а также офлайновых и онлайновых курсов, которые могут помочь овладеть SQL. Я рекомендую один из бесплатных онлайновых курсов W3Schools' SQL Tutorial[1], так как там пользователь имеет возможность составлять запросы прямо в браузере. Другой подход к обучению заключается в установке базы данных на компьютер пользователя. Установка и конфигурация основных баз данных, таких как MySQL и PostgreSQL, может оказаться делом непростым. Так что я настоятельно рекомендую начать с SQLite[2]: многие приложения в вашем смартфоне используют SQLite для хранения данных. Эта база данных бесплатная, простая в установке, сохраняет данные в единый переносимый файл, с ней вы быстро научитесь составлять SQL-запросы.

Если вы переживаете, что это старая технология, которую скоро затмят новинки, в исследовании O'Reilly 2014 Data Science Salary Survey Кинг и Маголас отмечают: «SQL был самым распространенным инструментом… Даже с бурным развитием технологий по работе с данными нет никаких признаков того, что SQL начинает сдавать позиции».

ПРОВЕРКА ФАЙЛА И ОПЕРАЦИИ С НИМ

В случаях, когда команде аналитиков приходится работать с большим количеством файлов с сырыми данными или с файлами большого

[1] URL: https://www.w3schools.com/sql/.

[2] Начать знакомство с SQL можно, например, с книги Дж. Крибича Using SQLite (O'Reilly).

объема, кто-то — необязательно все, поскольку аналитика все-таки командный спорт, — должен обладать элементарными знаниями Unix для проверки файлов и проведения операций с ними. В качестве альтернативы можно выбрать какой-нибудь из языков программирования, например Python, способный обеспечить эти функции и многие другие. Подробнее об этом в главе 5.

ПРИМЕР ЕЩЕ ОДНОГО ИНСТРУМЕНТА: ПОДСЧЕТ СТРОК ПРИ ПОМОЩИ *NIX-УТИЛИТЫ WC

Если вы знакомы со стандартными командами ОС *nix (то есть Unix и Linux), то можете пропустить эту часть. Всем остальным эта информация может оказаться полезной.

Предположим, вы получили данные в формате CSV-файла объемом 10 MB и вам нужно знать общее количество записей. Как их подсчитать? Открыть файл в Excel, пролистать до конца или воспользоваться комбинацией клавиш CTRL+↓ и посмотреть номер последней строки? Да, можно и так. А что, если файл будет объемом 100 MB? Конечно, Excel справится и с ним, но на выполнение этой задачи может уйти до десяти минут. Ладно, а как насчет файла объемом 1 GB? Здесь такой подход уже не сработает.

Ок, немного изменим условия задачи: теперь вы имеете дело с тремя CSV-файлами объемом 10 MB. Открыть каждый из них по отдельности в Excel? Допустим. А если у вас 300 таких файлов? Да, здесь явно нужен другой подход.

А что, если я скажу, что на решение этой задачи потребуется всего несколько секунд? Пакет стандартных команд ОС *nix представляет собой набор небольших специализированных утилит, обеспечивающих выполнение одной конкретной функции. **wc** представляет собой Unix-утилиту, выводящую количество слов (**w**ord **c**ount), а также строк и символов.

В: Но у меня нет доступа к *nix! У меня ОС Windows.

О: Ничего страшного, просто установите бесплатно cygwin[1]. Это позволит вам пользоваться командами Unix в ОС Windows.

В: Но у меня нет доступа к *nix! У меня OS X.

О: Mac OS X принадлежит семейству операционных систем Unix. Так что ваша цепочка действий следующая: идете в приложения Applications, открываете утилиты Utilities и кликаете на Terminal. Тадам! Можете пользоваться командами Unix.

[1] URL: https://www.cygwin.com/.

Формат команды элементарный: *wc -l filename*

`wc` — утилита для вывода количества слов, –l (символ) обозначает, что требуется вывести количество строк, а не слов, filename — название файла. Например:

```
$ wc -l weblog_20150302.log
1704190 weblog_20150302.log
```

($ — это подсказка или напоминание; у вас она может быть другой).

Этот пример показывает, что в файле *weblog* 1,7 млн строк. Для подсчета строк в каждом файле директории укажите название папки вместо имени файла:

```
wc -l mydatafiles/
123 file1.csv
456 file2.csv
579 total
```

Все очень просто. Утилита даже вывела итоговую строку. Я постоянно пользуюсь этой командой при проверке качества данных, чтобы оценить, сколько времени может занять загрузка набора данных в базу данных, а также для проверки, что все данные загрузились полностью.

Надеюсь, вы уловили главное: простые утилиты, научиться пользоваться которыми можно за несколько минут, способны значительно усилить набор аналитических навыков и повысить продуктивность работы.

Каким инструментом или утилитой научиться пользоваться, зависит от того, каким набором навыков вы уже владеете и какие у вас слабые места.

Будьте уверены, слабые места есть у всех. Последуйте моей рекомендации.

Если вам нужен дополнительный стимул, задумайтесь о следующем. В опросе на тему размера оплаты труда специалистов по работе с данными O'Reilly's 2013 Data Science Salary Survey приняли участие посетители двух крупных конференций Strata в 2012 и 2013 годах, при этом выяснилось следующее: размер оплаты труда положительно коррелировал с количеством инструментов, которыми пользовались респонденты.

В среднем респонденты использовали в работе 10 инструментов и их медианный доход составлял 100 тыс. долл. У тех, кто использовал 15 и более инструментов, показатель медианного дохода был 130 тыс. долл.

Еще более очевидно это отражено в опросе 2014 года[1] (рис. 4.4).

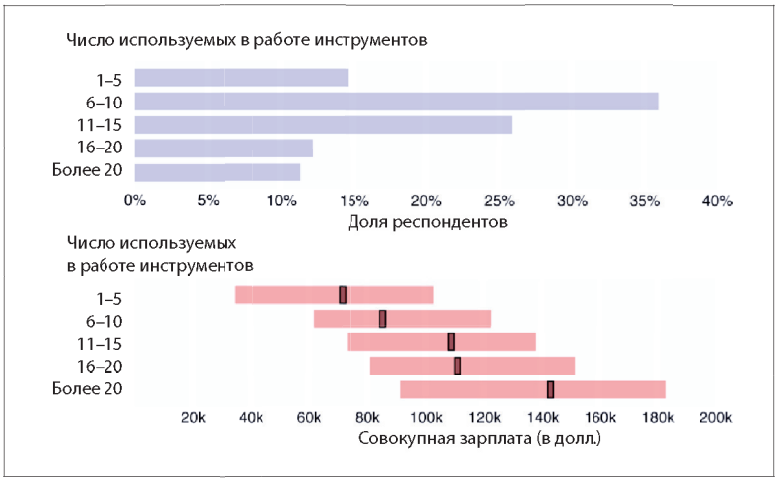


Рис. 4.4. Корреляция между применением разного числа инструментов и оплатой труда специалистов по работе с данными

Источник: опрос 2014 O'Reilly Data Science Salary Survey, рис. 1.13

В 2013 году авторы опроса сделали заключение:

> Есть веские основания утверждать, что владение такими инструментами, как R, Python, инструментарием Hadoop, D3, а также масштабируемыми инструментами машинного обучения, свидетельствует о более высокой квалификации аналитика, позволяя ему претендовать на более высокооплачиваемую позицию, чем когда аналитик владеет такими инструментами, как SQL, Excel и платформы RDB [реляционных баз данных]. Мы также пришли к выводу, что чем бóльшим числом инструментов способен пользоваться аналитик, тем лучше: если вы задумываетесь о том, чтобы научиться применять инструмент из набора Hadoop, лучше изучите сразу несколько.

Наконец, опрос 2014 года показал разницу в оплате труда почти в 15 тыс. долл. между аналитиками, умеющими работать с программным кодом, и не умеющими. Так что если это ваше слабое место, окажите себе услугу, научитесь программировать!

[1] URL: http://www.oreilly.com/data/free/files/stratasurvey.pdf.

Организация работы аналитиков в компании

Теперь, когда мы рассмотрели типы специалистов по аналитике и их навыки, можно перейти к вопросу организации их работы в контексте компании. Сначала давайте остановимся на двух крайних ситуациях.

ЦЕНТРАЛИЗОВАННАЯ МОДЕЛЬ

Есть центральная команда аналитиков, и все аналитики подотчетны ей. В этом варианте есть много преимуществ. Во-первых, команда может стандартизировать навыки, процесс обучения и применяемый инструментарий, кроме того, аналитики совместно используют ресурсы, что ведет к снижению расходов на приобретение лицензий на ПО. Во-вторых, команде аналитиков бывает легче продвигать результаты аналитической работы в компании. В-третьих, аналитики имеют возможность профессионального и личного общения, они могут чему-то научиться у коллег и поделиться с ними своим опытом. К тому же они ощущают себя частью команды единомышленников. В-четвертых, у них есть или может возникнуть ощущение большей объективности, поскольку успех их работы, как правило, не соотносится с успехом проектов, анализом которых они занимаются. Наконец, они способны продвигать основные источники данных в качестве единственных источников верных данных. Из недостатков этого способа организации работы аналитиков можно выделить то, что они оказываются в некоторой степени удалены от руководителей бизнеса и их целей, в результате чего стиль их работы может стать более бюрократическим[1]. Как отмечает Пиянка Джейн, «все должно подчиняться единому процессу, должны быть расставлены приоритеты и распределены ресурсы»[2].

ДЕЦЕНТРАЛИЗОВАННАЯ МОДЕЛЬ

При децентрализованной организации работы специалисты по анализу данных работают в отдельных подразделениях. Эти аналитики готовят отчеты для своих команд и разделяют их цели и задачи. Иными словами, их цели, отчеты и показатели — это цели, отчеты и показатели

---

[1] Rudin K. Big Impact from Big Data, 29 октября 2013 года, видеоклип, YouTube. URL: https://www.youtube.com/watch?v=RJFwsZwTBgg. Davenport T. H. and Harris J. G.. Analytics at Work. Boston: Harvard Business Press, 2007.

[2] Jain P. To Centralize Analytics or Not, That is the Question, Forbes, February 15, 2013. URL: https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/piyankajain/2013/02/15/to-centralize-analytics-or-not/&refURL=&referrer=.

подразделения, в котором работает аналитик. Минус этого подхода в том, что аналитик оказывается оторванным от других аналитиков компании. Это приводит к риску избыточных усилий, несовпадения инструментария, навыков, определений показателей и реализации. У аналитиков из разных команд меньше возможность общения и обмена профессиональным опытом. Децентрализованная модель наиболее распространена, ее придерживаются 42% респондентов нашего опроса. По Дэвенпорту и др. (с. 108), это фактор, отражающий «незрелость аналитики». Авторы не поясняют свою позицию, но моя интерпретация заключается в том, что довольно сложно демонстрировать качественные результаты на более высоком уровне аналитической работы, например как в отделе исследования операций, где занимаются оптимизацией или проблемами прогнозирования, без централизованной координации усилий, практического опыта и контроля.

У каждой из этих моделей есть свои плюсы и минусы (они перечислены в табл. 4.2). В первом случае аналитик в большей мере ощущает поддержку, имеет возможность профессионального общения и обмена опытом, у него более четкий карьерный путь. Во втором случае распределение ресурсов зависит от политики руководителя, но предположительно уменьшается срок выполнения работы.

Таблица 4.2. Преимущества централизованной модели организации работы аналитиков над децентрализованной моделью. (Недостатки выступают оборотной стороной преимуществ в любом из столбцов.) Повышение уровня профессионализма может происходить в обоих случаях (см. объяснение в тексте)

Преимущества	**Централизованная модель**	**Децентрализованная модель**
Четкий карьерный путь	+	
Прямой доступ в любое время		+
Более короткий срок выполнения работы		+
Концентрация профессиональных знаний и опыта	+	
Стандартизированный инструментарий и процесс обучения	+	
Стандартизированные показатели	+	
Меньше бюрократии		+
(Воспринимаемая) объективность	+	
Более высокий уровень профессиональных знаний и навыков	?	?

Организации, находящиеся на преобразованном уровне, на 63% чаще, чем организации на желательном уровне (см. главу 1), «используют централизованное подразделение как основной источник аналитики». Однако здесь в действие вступают искажающие факторы (в частности, величина компании и общее количество специалистов по анализу), так как в компаниях на преобразованном уровне аналитики также работают в бизнес-подразделениях[1].

Логично предположить, что при децентрализованной модели у аналитиков сильнее повышается уровень профессиональных знаний, например, у них формируется более глубокое понимание данных по клиентам, аналитических процессов и показателей. К сожалению, при таком уровне экспертных знаний повышается риск для компании в целом, если эти несколько высококлассных специалистов ее покинут. (При централизованной модели более высока вероятность избыточности знаний, так как аналитики переключаются между разными направлениями бизнеса.) Это может означать, что уровень профессиональных знаний в среднем фактически *ниже* при децентрализованной модели, если аналитики часто увольняются, а на их место приходят новички, на обучение которых требуются годы.

Джеб Стоун[2] считает, что при централизованной модели с несколькими стандартными технологиями:

> …чтобы повысить ценность для организации, аналитик должен овладеть этими дополнительными технологиями, обучиться этим смежным специализированным направлениям бизнеса и приблизиться к тому уровню и качеству работы, которые задают старшие аналитики. Без четко обозначенного карьерного пути у аналитиков может оказаться велик соблазн обучиться новым навыкам за счет компании, вне зависимости от того, насколько это ей нужно, а затем перейти к тому работодателю, который будет ему больше платить за эти навыки. И есть еще один аспект: ведущие аналитики, скорее всего, будут избегать компаний с децентрализованной моделью организации аналитической работы, поскольку они знают, что у них уйдет гораздо больше времени на продвижение по карьерной лестнице. К тому же

[1] LaValle S., Hopkins M. S., Lesser E., Shockley R. and Kruschwitz N. Analytics: the New Path to Value, MIT Sloan Management Review 52, no. 2 (2010): Figure 9. URL: http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/.

[2] Stone J. Centralized vs Decentralized Analytics: All You Need To Know, April 22, 2012. URL: http://jebstone.com/2012/04/centralized-vs-decentralized-analytics-all-you-need-to-know/.

в подобной компании вряд ли будет стимулирующая программа, адекватная их профессиональным достижениям.

В попытках создать структуру, максимально сохраняющую преимущества и минимизирующую недостатки, возникла так называемая смешанная модель — подобная используется в компании Facebook. В ее рамках присутствует центральная команда аналитиков, и таким образом стандартизированы инструментарий, процесс обучения и другие профессиональные аспекты. При этом физически специалисты по работе с данными находятся в разных бизнес-подразделениях и разделяют их цели. Таким образом компании удается извлекать преимущества из тесного взаимодействия разных сотрудников и наличия аналитических стандартов. К недостаткам этой модели можно отнести то, что возникает ситуация, когда аналитикам может быть необходимо отчитываться перед несколькими руководителями: по линии аналитической работы и конкретного бизнес-направления. Это может вести к риску возникновения конфликтов или получения противоречивых посылов.

При децентрализованной модели организации аналитической работы могут потребоваться способы объединения аналитиков, чтобы они могли обмениваться опытом и профессиональными навыками, посещать обучающие мероприятия, обсуждать источники данных, показатели, результаты проведенного анализа. Один из подходов — и именно его мы применяем в компании Warby Parker — заключается в создании гильдии аналитиков, «организованной группы людей, объединенных общим профессиональным или иным интересом». Это позволяет аналитикам из разных подразделений, а в нашем случае из разных зданий, общаться и обсуждать разные вопросы. Кроме того, моя команда специалистов по работе с данными получает возможность проводить обучение инструментам бизнес-аналитики и статистики.

Подобная гильдия напоминает матричную структуру, и для ее создания и функционирования требуется серьезная поддержка со стороны руководителей или начальников подразделений, которым подчиняются эти аналитики, а также со стороны руководителей более высокого уровня. Аналитики должны заручиться согласием своих руководителей на то, что им будут выделять время на участие в гильдии.

Другие виды организационных структур[1], более характерные для крупных компаний, перечислены ниже.

[1] Davenport T. H. and Harris J. G. Analytics at Work. Boston: Harvard Business Press, 2007. Khalil E. and Wood K. Aligning Data Science — Making Organizational Structures Work, (Tysons Corner, VA: Booz Allen Hamilton, Inc., 2014).

Консалтинговая структура

В некоторых компаниях централизованная модель модифицирована таким образом, что аналитиков нанимают в подразделения в формате консалтинговой структуры. При слабой исполнительной власти есть риск, что аналитик соблазнится на деньги или поддержит более убедительного руководителя, но при этом для компании его работа не будет иметь большой ценности.

Функциональная структура

Форма централизованной модели, при которой команда аналитиков включена в функциональное бизнес-подразделение и в основном «работает» на него. При этом при необходимости она может решать задачи других подразделений компании. В некоторых случаях вся команда аналитиков может даже перейти в другое подразделение.

Центр передового опыта

Несколько напоминает смешанную структуру, но в большем масштабе, кроме того, ряд аналитических специалистов, таких как специалисты по статистике, остается в «центральном узле». Таким образом, аналитическая работа проводится как в отдельных подразделениях, так и центральной командой специалистов.

В табл. 4.3 перечислены разные организационные структуры и приведены примеры компаний каждого типа. Тем не менее стоит подчеркнуть, что это *идеализированные* структуры: на практике границы между ними часто размыты, и образуются разные смешанные типы. Например, в компании Warby Parker применяется децентрализованная модель, в которой аналитики отчитываются только перед руководителем по конкретному бизнес-направлению, при этом присутствуют элементы модели центра передового опыта, так как в компании есть центральная команда специалистов по аналитической работе, которые обеспечивают поддержку с точки зрения углубленной аналитики (а также наличие инструментов бизнес-аналитики, обучение специалистов и стандарты деятельности). Однако ожидается, что эта структура будет меняться по мере «взросления» аналитики в организации.

Нет единого ответа на вопрос, какая структура лучше всех. Все зависит от размера компании и области, в которой она действует. Например,

не имеет смысла внедрять модель центра передового опыта, если в компании всего пять аналитиков. Она будет эффективна в организациях с числом сотрудников больше 25 тыс. человек. Определенная структура может адекватно отвечать задачам компании на данном этапе ее развития, но по мере роста компании может потребоваться реорганизация этой структуры.

Таблица 4.3. Примеры разных структур организации аналитической работы

Модель организационной структуры	**Аналитики отчитываются перед**		**Примеры**
	центральной аналитической командой	**руководителями бизнеса**	
Централизованная	+		Mars, Expedia, One Kings Lane
Децентрализованная		+	PBS, Dallas Mavericks
Смешанная	+	+	Facebook, Ford, Booz Allen Hamilton
Функциональная структура	+		Fidelity
Консалтинговая структура	+		eBay, United Airlines
Центр передового опыта	+	+	Capital One, Bank of America

Тем не менее, опираясь на результаты ежегодного технологического исследования Accenture и анализ более 700 специалистов[1], Дэвенпорт и др. (с. 106) утверждают:

> Мы полагаем, что централизованная модель и модель центра передового опыта (или смешанные модели, включающие элементы обеих этих моделей) способны предложить самые существенные потенциальные преимущества тем

[1] Harris J. G., Craig E. and Egan H. How to Organize Your Analytical Talent (Dublin: Accenture Institute for High Performance, 2009).

компаниям, которые готовы предпринять корпоративный подход к аналитике. У аналитиков, работающих в рамках этих моделей, значительно выше уровень вовлеченности, удовлетворенности работой, воспринимаемой поддержки со стороны компании, ресурсов и лояльности по отношению к компании[1].

В главе 11 мы обсудим, какое место занимают эти команды в разрезе всей структуры компании в целом и кому из топ-менеджеров компании подчиняются. Однако до этого давайте подробнее изучим то, чем занимаются аналитики, — процесс анализа.

[1] Davenport T. H., Harris J. G. and Morison R. Competing on Analytics. Boston: Harvard Business Press, 2010.

ГЛАВА 5
Анализ данных

Если достаточно долго мучить данные, они признаются [в чем угодно].
Рональд Коуз[1]

Следующие три главы посвящены сути аналитической работы: непосредственно анализу данных, целям анализа с позиции компании и тому, как проводить *результативный* анализ данных.

Мы рассмотрим такие аспекты, как виды анализа данных, разработка показателей, извлечение практических выводов, презентация этих выводов, идей и рекомендаций руководителям. В главе 6 мы обсудим разработку показателей и ключевых показателей эффективности деятельности (KPI), а глава 7 посвящена визуализации данных и сторителлингу[2]. В этой главе, первой из трех, речь пойдет непосредственно об анализе данных.

Важно отметить, что мы не будем говорить о том, *как* проводить анализ или статистическое исследование, — на эту тему есть много других более полных источников (см. список дополнительной литературы). Мы сосредоточимся на цели анализа данных: что это означает? К какому результату стремятся аналитики? Какие инструменты входят в их профессиональный набор? Мы вернемся к идее разных уровней аналитики, о которой уже упоминалось в главе 1, и изучим другие точки зрения на виды аналитики.

Наша цель — выделить ряд инструментов статистики и визуализации, которые аналитики могут использовать в своей работе. Дополнительная

[1] Рональд Коуз (1910–2013) — американский экономист, лауреат Нобелевской премии по экономике. *Прим. перев.*

[2] Сторителлинг (от *англ.* storytelling) — маркетинговый прием, использующий медиапотенциал с целью передачи информации и транслирование смыслов посредством рассказывания историй. *Прим. перев.*

цель заключается в том, чтобы стимулировать их применять подходящие инструменты, а при необходимости изучить более сложные инструменты, способные обеспечить более глубокий уровень понимания конкретной проблемы.

Для изготовления деревянного стола опытному столяру требуется качественный исходный материал: древесина красного дерева, набор столярных инструментов, например стамеска и угольник, и профессиональные знания, когда и как пользоваться этими инструментами. Отсутствие хотя бы одного из трех компонентов заметно скажется на качестве конечного продукта. То же самое касается и аналитической работы. Для производства аналитического продукта, имеющего реальную ценность, не обойтись без исходного материала в виде качественных данных, инструментария в формате различных аналитических методов и техник, а также профессиональных знаний, когда и как пользоваться всеми этими инструментами для решения задачи.

Что такое анализ данных?

Уделим немного времени самому термину «анализ». Он происходит от древнегреческого ἀνά [ana] + λύω [luō], что означает «освобождать», «распутывать». В этом есть смысл, но слишком высокопарный, чтобы помочь нам уловить, что это действительно означает. Для целей бизнеса можно воспользоваться определением Марио Фариа из главы 1:

> *Анализ — преобразование данных в выводы, на основе которых будут приниматься решения и строиться действия с помощью людей, процессов и технологий.*

Давайте остановимся на этом подробнее. Надеюсь, из глав 2 и 3 у вас уже сложилось понимание, что такое массив данных, а вот что такое аналитические выводы?

Согласно «Википедии», аналитические выводы — понимание конкретных причин и следствий в конкретном контексте[1]. В английском языке у этого термина (insight) есть несколько сопутствующих значений:

— информация;

— «озарение» — понимание внутренней сути вещей и процессов;

— самоанализ;

[1] URL: https://en.wikipedia.org/wiki/Insight.

— проницательность, способность делать глубокие наблюдения и выводы;

— понимание причин и следствий на основе установления взаимосвязи и поведения в рамках модели, контекста или сценария.

Итак, понимание взаимосвязи причин и следствий, понимание внутренней природы вещей и процессов и так далее. Это будет нам полезно.

Термин «информация»[1], то есть «результат обработки данных для придания им контекста и смысла», часто используется как синоним термина «данные», хотя технически это не одно и то же (см. ниже текст в рамке, а также статью The Differences Between Data, Information and Knowledge («Разница между понятиями “информация”, “данные” и “знания”»)[2].

ДАННЫЕ, ИНФОРМАЦИЯ И ЗНАНИЯ

Данные представляют собой сырые, необработанные факты об окружающем мире. Информация — собранные, обработанные данные, в то время как знания — это набор ментальных моделей и убеждений об окружающем мире, который сформировался на основе информации, полученной на протяжении какого-то периода времени.

Температура на данный момент составляет 6 °C. Это количественный факт. Он существует и соответствует действительности вне зависимости от того, зафиксировал ли его кто-то. К сожалению, этот факт бесполезен (для всех, кроме меня), так как из-за отсутствия контекста (когда? где?) он не позволяет сделать никаких выводов.

В Нью-Йорке 2 ноября 2014 года в 10 утра температура составила 6 °C. У этих данных есть контекст. Однако это по-прежнему лишь констатация факта без интерпретации.

Температура 6 °C гораздо ниже климатической нормы. Это информация. Мы обработали данные и объединили их с другими данными, чтобы определить понятие климатической нормы и оценить, как соотносятся значения.

При температуре 6 °C на улице прохладно, я надену пальто. Вы объединили информацию за какой-то период времени и построили мыслительную модель, что это означает. Это знания. Конечно, все эти модели относительны. Например, житель Аляски может посчитать температуру 6 °C в ноябре не по сезону теплой.

[1] URL: http://foldoc.org/information.

[2] URL: http://www.infogineering.net/data-information-knowledge.htm.

Исходя из глубины информации, мы вновь можем вернуться к подробному определению анализа (рис. 5.1). Хотя в нем по-прежнему остаются такие термины, как «понимание» и «контекст», надеюсь, теперь у вас более четкое представление о том, что такое анализ, по крайней мере концептуально. На этом новом уровне понимания давайте изучим набор инструментов, находящийся в распоряжении аналитиков. Сейчас речь идет не о программных инструментах, таких как Excel или R, а о статистических инструментах и о *видах* анализа данных, которые можно проводить.

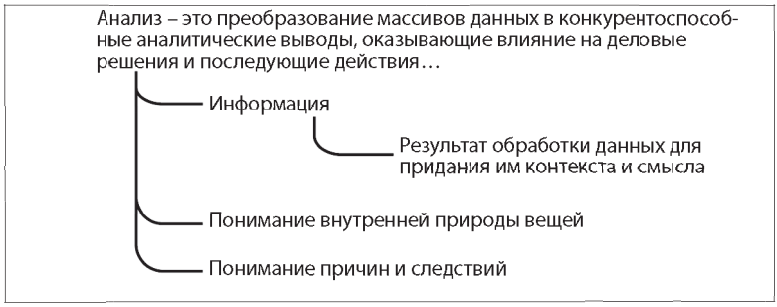

Анализ – это преобразование массивов данных в конкурентоспособные аналитические выводы, оказывающие влияние на деловые решения и последующие действия…
Информация
Результат обработки данных для придания им контекста и смысла
Понимание внутренней природы вещей
Понимание причин и следствий

Рис. 5.1. Результат двухуровневого раскладывания определения термина «анализ»

Виды анализа данных

Джеффри Лик, старший преподаватель биостатистики в Университете Джонса Хопкинса, а также один из редакторов блога о статистике[1], выделяет шесть типов анализа данных[2]. Они перечислены далее от простого к сложному:

— описательный (descriptive);

— разведочный (exploratory);

— индуктивный (inferential);

— прогностический (predictive);

— каузальный (причинно-следственный) (causal);

— механистический (mechanistic).

[1] URL: http://simplystatistics.org/.

[2] По крайней мере, он рассматривает эти шесть типов анализа данных в рамках своего курса Data Analysis Course.

Мы рассмотрим первые пять типов анализа. Механистический тип в большей степени связан с фундаментальной наукой, исследованиями и разработками, и к нему больше подходит термин «моделирование», чем «анализ». Механистическое моделирование и анализ отличаются очень глубоким пониманием системы, которое приходит в результате многолетнего контролируемого изучения стабильной системы посредством большого числа экспериментов. Именно на этом основана моя ассоциация с фундаментальной наукой. Это редкость для большинства компаний, за некоторыми исключениями, такими как научно-исследовательские подразделения фармацевтических компаний и инженерно-проектные подразделения технических компаний. Иными словами, если вы проводите анализ данных на этом уровне, который представляет собой вершину анализа, то практически наверняка вам не требуется читать в этой книге, как его выполнять. Если вернуться к главе 1, то сейчас у вас должен прозвучать звоночек. Ранее мы говорили о восьми уровнях *аналитики*. Сейчас мы говорим о шести типах *анализа* данных, при этом у нас встретилось всего одно общее слово — «прогностический». Что все это значит?

В предыдущем списке перечислены типы статистического анализа. Важно отметить, что они могут относиться к разным уровням аналитики. Например, на основе разведочного анализа данных (о котором шла речь в главе 2) можно подготовить ad hoc отчет (уровень аналитики 2). Также на его основе можно сформулировать бизнес-логику для системы оповещения (уровень аналитики 4), например определить 98-й процентиль в распределении и установить сигнал оповещения, если соответствующий показатель превысит этот уровень.

На рис. 5.2 показана попытка соотнести эти два списка: уровни аналитики (по вертикали) и пять типов анализа данных (по горизонтали). Интенсивность цвета каждой ячейки обозначает примерную оценку усилий или времени, затраченных на проведение этого типа анализа. Например, подготовка стандартных отчетов обычно осуществляется на основе описательного и разведочного типов анализа, при этом крайне маловероятно использование причинно-следственных моделей. С другой стороны, аналитика оптимизации строится на описательном и разведочном анализе, но в первую очередь сосредоточена на прогностическом и, возможно, причинно-следственном анализе.

Необходимо прояснить один момент. Существует множество других видов количественного анализа, например анализ выживаемости, анализ социальных сетей, анализ временны́х рядов. При этом каждый из них связан с конкретной областью профессиональных знаний или типом данных, а применяемые аналитические инструменты и подходы *включают*

в себя шесть более общих аналитических инструментов и подходов. Например, при анализе на основе временны́х рядов можно вычислить период действия явления (описательный анализ), затем определить переменную во времени (разведочный анализ) и, наконец, смоделировать и прогнозировать будущие показатели (прогностический анализ). Вы получаете общую картину. Иными словами, перечисленные шесть классов представляют собой архетипы анализа. Кроме того, есть другие типы качественного анализа. Например, анализ основных причин, метод «Пять “почему”» от Toyota[1] и методология «Шесть сигм». Принимая это во внимание, давайте рассмотрим пять типов анализа.

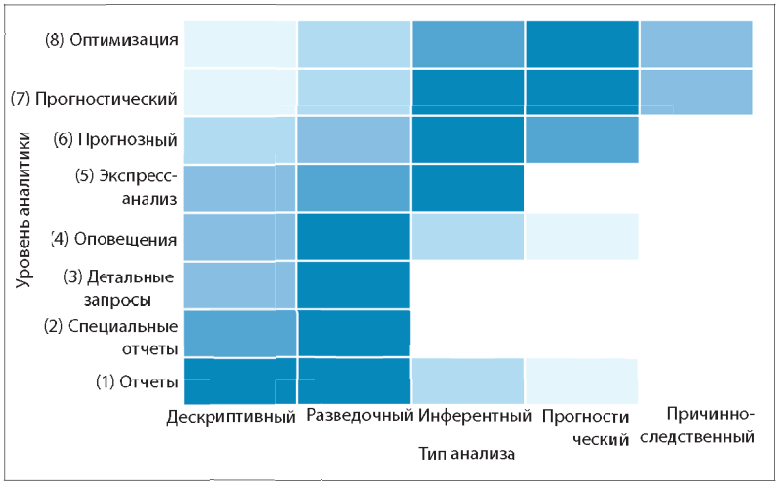


Рис. 5.2. Примерное соотношение между уровнем аналитики (по вертикали) и типом анализа (по горизонтали). Объяснение см. в тексте

СЛОВАРЬ ТЕРМИНОВ

Вы еще не запутались во всех этих «показателях», «переменных», «значениях»? Не переживайте. Эти термины пересекаются, и насчет их определении нет согласия. Ниже представлены мои варианты.

Переменная (Variable)

Показатель, который склонен меняться со временем, пространством или единицами выборки. Например, «Допустим, переменная v = скорость движения автомобиля» или «Пол — категориальная переменная».

[1] URL: https://en.wikipedia.org/wiki/5_Whys.

Измерение (Dimension)

Это переменная. В то время как термин «переменная» чаще используют ученые и программисты, для представителей деловых кругов больше характерно употребление термина «измерение». Измерение — переменная, характеризующая факты и количественные показатели, она может отражать параметр категории или времени, а также рейтинга, рэнкинга или числа. Например, вы можете проанализировать совокупный объем продаж (значение) относительно страны (измерение) или года (измерение) или же рассчитать процент отказов (значение) относительно пола (измерение). В моем представлении измерения, как правило, находятся на оси *x*, а показатели — на оси *y*.

Значение (Measure)

Количественный показатель какого-либо свойства объекта, например длина, или стандартная единица измерения. В области бизнес-аналитики этот термин обычно относится к функции (например, BMI) или агрегированному значению, например минимальное, суммарное или среднее значение количественных данных. Может рассматриваться в виде чистого или производного значения чего-либо.

Показатель (Metric)

Функция от двух или более значений (с точки зрения измерения) или просто значение (в функциональном смысле). Производное значение.

Статистический показатель (Statistic)

Определенный показатель какого-то свойства в выборке значений, например среднее арифметическое = 6,3. Это функция, примененная к набору числовых данных, которая представляет собой отдельное значение. Несколько сбивает с толку, что и сама функция, и итоговое ее значение — статистические показатели.

Ключевые показатели эффективности деятельности (Key performance indicator)

В контексте ведения бизнеса этот показатель связан с целью деятельности и/или некоторыми основными ценностями (подробнее о KPI мы поговорим в следующей главе). То есть этот показатель связан с целью бизнеса или стартовой точкой.

ОПИСАТЕЛЬНЫЙ АНАЛИЗ

Наиболее простой тип анализа данных — описательный (дескриптивный). Он обеспечивает количественное описание набора данных. Важно отметить, что этот тип анализа касается только выборки данных, по которой проводится анализ, и не описывает ту совокупность, из которой он взят. На основании описательного анализа часто формируются данные, которые отображаются в дашбордах, например количество новых пользователей за неделю или размещенных заказов с начала года (см. раздел «Дашборды» в главе 7).

Давайте начнем с одномерного анализа, то есть описывающего одну переменную (ряд или поле) из набора данных. В главе 2 мы уже обсуждали составление пятичисловой сводки, однако есть множество других возможных статистических показателей; их можно условно разделить на меры среднего уровня («середина» данных), меры рассеивания (разброса данных) и формы распределения. Ниже перечислены **показатели, относящиеся к числу простейших**, но при этом наиболее важных.

Размер выборки

Количество единиц (записей) в выборке данных.

Далее перечислены **меры среднего уровня**.

Среднее значение

Чтобы найти среднее арифметическое, нужно сложить все значения и разделить на их количество.

Среднее геометрическое

Этот показатель применяется для определения среднего значения при наличии мультипликативного эффекта, например сложных процентов со ставкой, меняющейся из года в год. Чтобы найти среднее геометрическое, нужно перемножить все значения и извлечь из них корень. Степень корня определяется количеством значений. Если вы получили 8% в первый год, а затем по 6% следующие три года, средняя процентная ставка составит 6,5%.

Среднее гармоническое

Средним гармоническим называется число, обратное среднему арифметическому их обратных. Например, если вы доехали

до магазина со скоростью движения 80 км/ч, а на обратной дороге попали в пробку и скорость вашего движения составила 32 км/ч, ваша средняя скорость составит не 56, а 47 км/ч.

Медиана

Медиана — 50-й процентиль.

Мода

Наиболее часто встречающееся значение.

К **мерам рассеяния** относятся следующие.

Минимум

Наименьшее значение в выборке (0-й процентиль).

Q1

25-й процентиль. Значение выборки такое, что одна четвертая остальных значений выборки меньше него.

Q3

75-й процентиль. Значение выборки такое, что одна четвертая остальных значений выборки больше него.

Максимум

Максимальное значение в выборке (100-й процентиль).

Межквартильный размах

Центральные 50% данных, разность между третьим и первым квартилями.

Размах

Разница между максимумом и минимумом.

Стандартное отклонение

Наиболее распространенный показатель рассеивания значений случайной величины относительно ее математического ожидания. Вычисляется как квадратный корень из дисперсии. Измеряется в тех же единицах, что и сама случайная величина.

Дисперсия

Мера разброса значений случайной величины относительно ее математического ожидания. Вычисляется возведением стандартного отклонения в квадрат. Измеряется в квадратах единицы измерения случайной величины.

Стандартная ошибка

Вычисляется путем деления стандартного отклонения на квадратный корень размера выборки. Показывает ожидаемое стандартное отклонение среднего значения выборки, если бы мы повторно получали выборки такого же размера из того же источника генеральной совокупности.

Коэффициент Джини

Количественный показатель, изначально разработанный, чтобы показать степень неравенства при распределении доходов. Тем не менее его можно использовать более широко. Он равен половине ожидаемой абсолютной разницы между доходами двух случайно выбранных людей, деленной на средний доход.

Меры формы включают следующие.

Коэффициент асимметрии

Величина, характеризующая асимметрию распределения. Коэффициент асимметрии положителен, если правый хвост распределения длиннее левого, и отрицателен в противном случае. Число фолловеров среди пользователей сервиса Twitter характеризуется положительным коэффициентом асимметрии (см., например, отчет An In-Depth Look at the 5% of Most Active Users[1] и статью Tweets loud and quiet[2]).

Коэффициент эксцесса

Мера остроты пика распределения случайной величины. У распределения с высоким коэффициентом эксцесса[3] острый пик и плоские хвосты. На это стоит обратить внимание при инвестировании, так как это означает вероятность более резких колебаний по сравнению с переменной с нормальным распределением.

---

[1] URL: https://www.sysomos.com/2009/08/05/exploring-twitters-most-active-users/.

[2] URL: https://www.oreilly.com/ideas/tweets-loud-and-quiet.

[3] URL: https://en.wikipedia.org/wiki/Kurtosis.

Кроме того, мне кажется, что тип распределения также можно назвать полезной описательной статистикой. Например, нормальное распределение (распределение Гаусса), логарифмически нормальное распределение, экспоненциальное распределение и унимодальное распределение — обычные. Зная тип, а следовательно, и форму распределения, можно узнать его потенциальные характеристики (например, что в нем могут быть редкие, но сильно отклоняющиеся значения), понять логику процесса генерации данных, а также определить, какие еще показатели требуется собрать. Например, если распределение представляет собой ту или иную форму экспоненциального закона, как распределение фолловеров в Twitter, очевидно, что следует вычислить отрицательный показатель экспоненты, который представляет собой важный критерий.

Не все переменные — непрерывные. Например, пол и продуктовая линейка относятся к категориальным переменным. Таким образом, описательный анализ может включать таблицы частотности для разных категорий или факторные таблицы, подобные следующей.

Объем продаж по регионам					
Пол	**Западный**	**Южный**	**Центральный**	**Восточный**	**Итого**
Мужской	3485	1393	6371	11 435	22 684
Женский	6745	1546	8625	15 721	32 637
Итого	10 230	2939	14 996	27 156	**55 321**

На этом уровне анализа проводящий его специалист должен знать, по какому критерию следует группировать данные, и понимать, когда какие-то данные выделяются из общей массы и представляют интерес. Например, в предыдущей таблице интересно, почему настолько велика доля женщин, совершающих покупки, в западном регионе.

При работе с двумя переменными описательный анализ может включать меры ассоциации, например вычисление коэффициентов корреляции и ковариации.

Цель описательного анализа состоит в числовом описании основных характеристик выборки. Он должен прояснять основные значения, отражающие распределение данных, кроме того, он может описывать взаимоотношения между переменными с показателями, описывающими ассоциации, или в сводных таблицах.

Некоторые из этих простых показателей могут оказаться весьма ценными сами по себе. Возможно, вам потребуется узнать и отследить среднее число заказов или наибольшую длительность их выполнения для

разрешения практического вопроса с клиентом. Таким образом, этих данных может быть достаточно для составления стандартного и ad hoc отчетов, запроса или оповещения (уровни аналитики 1–4), и это может принести пользу компании. Кроме того, вы можете убедиться в качестве данных. Например, если максимальный возраст игрока, который зарегистрировался на сайте игры — «стрелялки» от первого лица, указан как 115 лет, то либо пользователь ошибся при вводе этой информации, либо в графе с датой рождения была установлена дата по умолчанию 1900 (ну, или это *реально* крутая бабушка). Помочь это определить могут простые минимум и максимум, размах выборки и гистограммы.

Наконец, описательный анализ обычно бывает первым шагом — возможностью познакомиться с данными — к более глубокому анализу.

РАЗВЕДОЧНЫЙ АНАЛИЗ

Описательный анализ — важный первый шаг. При этом просто итоговых цифр может быть недостаточно. Одна из проблем заключается в том, что большое число значений сводится к нескольким итоговым цифрам. А потому не стоит удивляться, что одни и те же итоговые статистические показатели могут описывать разные выборки с разным распределением данных, формами и свойствами.

На рис. 5.3 представлены две выборки с одинаковым средним значением, равным 100, но очень разным распределением.

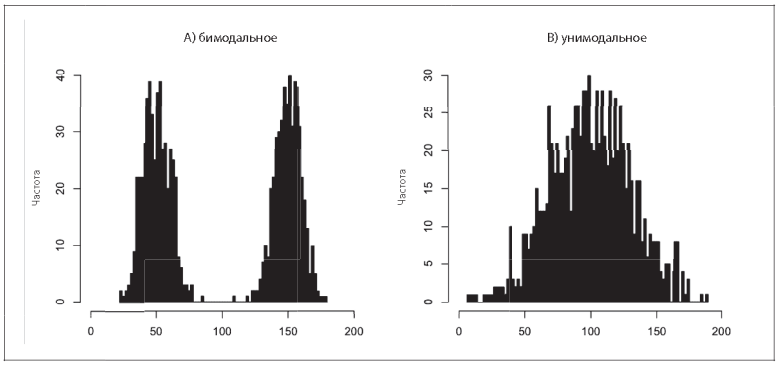


Рис. 5.3. A) бимодальное распределение и B) унимодальное распределение. В обоих случаях среднее значение одинаковое, примерно равно 100

Теперь это кажется не таким удивительным. У нас имеется простой итоговый статистический показатель — среднее значение одной

переменной. Существует множество потенциальных «решений», или выборок, которым может соответствовать это значение.

Сейчас я покажу вам гораздо более удивительный пример. Предположим, у вас четыре набора данных с двумя переменными со следующими характеристиками.

Характеристика	**Значение**
Размер выборки в каждом случае	11
Среднее значение переменной *x* в каждом случае	9
Дисперсия переменной *x* в каждом случае	11
Среднее значение переменной *y* в каждом случае	7,5
Дисперсия переменной *y* в каждом случае	4,122 или 4,127
Корреляция между *x* и *y* в каждом случае	0,816
Прямая линейной регрессии в каждом случае	$y = 3{,}00 + 0{,}500x$

Это система с жесткими заданными ограничениями. Значит, графики этих четырех наборов данных с идентичными статистическими характеристиками должны быть достаточно похожими, не так ли? А вот рис. 5.4 показывает, что это далеко не так.

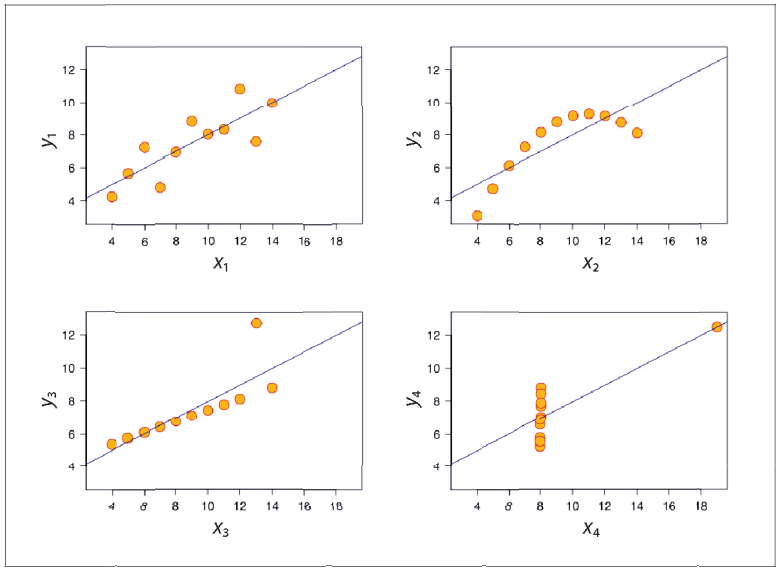


Рис. 5.4. Квартет Энскомба. В каждом из четырех наборов данных идентичны среднее значение x, среднее значение y, дисперсия x, дисперсия y, корреляция и прямая линейной регрессии (до двух знаков после запятой)

Источник: https://en.wikipedia.org/wiki/Anscombe's_quartet

Это так называемый квартет Энскомба[1], названный по имени математика и статистика Фрэнсиса Энскомба, который составил его в 1973 году. Энскомб выступил против существовавшей на тот момент доктрины в области статистических вычислений, которая гласила, что:

1) числовые данные точные, а графики — приблизительные;
2) для каждого конкретного вида статистических данных существует только один набор вычислений, обеспечивающий правильный статистический анализ;
3) выполнение сложных расчетов — единственно верный путь, изучение данных только вводит в заблуждение.

Энскомб утверждал:

> Большинство статистических вычислений строятся на предположениях относительно поведения данных. Эти предположения могут оказаться неверными, и тогда результаты вычислений тоже будут содержать ошибку. Всегда следует пытаться проверять, являются ли предположения верными. А если они ошибочны, мы должны быть способны понять, что с ними не так. В этом весьма полезны графики.

Применение графиков для визуализации и изучения данных получило название разведочного анализа данных. Наибольшую известность он приобрел благодаря продвижению американским математиком Джоном Тьюки в книге Exploratory Data Analysis (Pearson), опубликованной в 1977 году. При правильном подходе графики помогают видеть более масштабную картину, а также отмечать очевидные или необычные закономерности (это врожденное свойство человеческого мозга). Нередко аналитические выводы и понимание данных начинают формироваться именно на этом этапе. Почему у этой кривой такое отклонение? В какой момент наступает снижение возврата на маркетинговые расходы?

Разведочный анализ позволяет опровергнуть или подтвердить наши предположения относительно данных. Поэтому, когда в главе 2 шла речь о качестве данных, я рекомендовал использовать команду `pairs()` в среде R. Часто у нас сформированы обоснованные ожидания, что может быть не так с качеством данных, в отличие от ожиданий, какими должны быть достоверные данные.

[1] Anscombe F. J. Graphs in statistical analysis, American Statistician 27 (1973): 17–21.

По мере того как мы набираемся опыта и знаний в профессиональной области, у нас развивается интуитивное понимание, какие факторы и возможные отношения могут быть задействованы. Разведочный анализ, с его широким набором способов рассмотреть данные и их взаимоотношения, предлагает набор «луп» для изучения системы.

Это, в свою очередь, помогает специалисту по анализу данных выдвинуть новые гипотезы относительно того, что может произойти, если вы понимаете, какие переменные находятся под вашим контролем и какими рычагами вы можете воспользоваться для движения показателей, например выручки или конверсии, в нужном направлении. Кроме того, разведочный анализ способен показать пробелы в наших знаниях и определить, что можно сделать для их ликвидации.

Для одномерных непрерывных (действительные числа) или дискретных данных (целые числа) обычно строят диаграмму «стебель-листья» (рис. 5.5), гистограммы (рис. 5.6) и диаграммы размаха, или коробчатые диаграммы (рис. 5.7).

Стебель	Листья
1	0 3 6
2	1 6 7 8
3	5 5 6
4	1 1 5 6 9
5	0 3 6 8

Рис. 5.5. Диаграмма «стебель-листья»

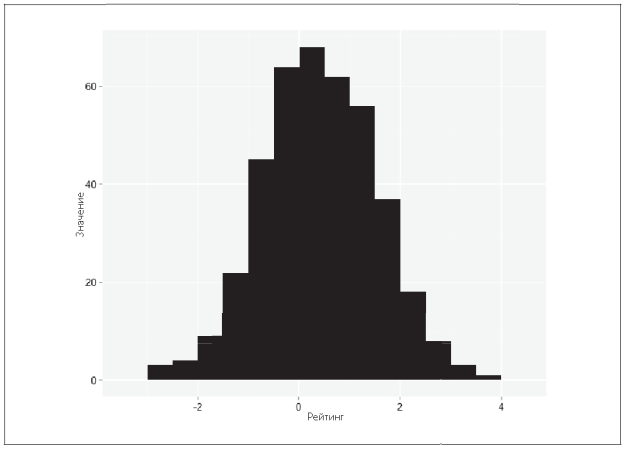


Рис. 5.6. Гистограмма

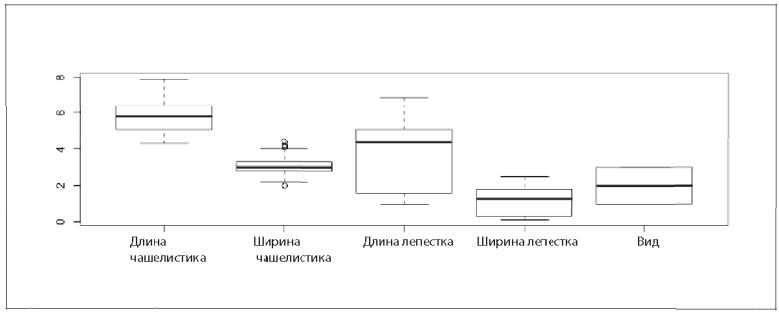


Рис. 5.7. Коробчатая диаграмма

Если гистограмма строится в таком масштабе, что ее площадь равна 1, это функция плотности распределения вероятностей.

Еще один полезный способ представить те же самые данные — составить интегральную функцию распределения.

Это может выделить интересные точки распределения, включая основные опорные точки.

На рис. 5.8, 5.9, 5.10 представлены основные графики для одномерных категориальных (качественных) переменных.

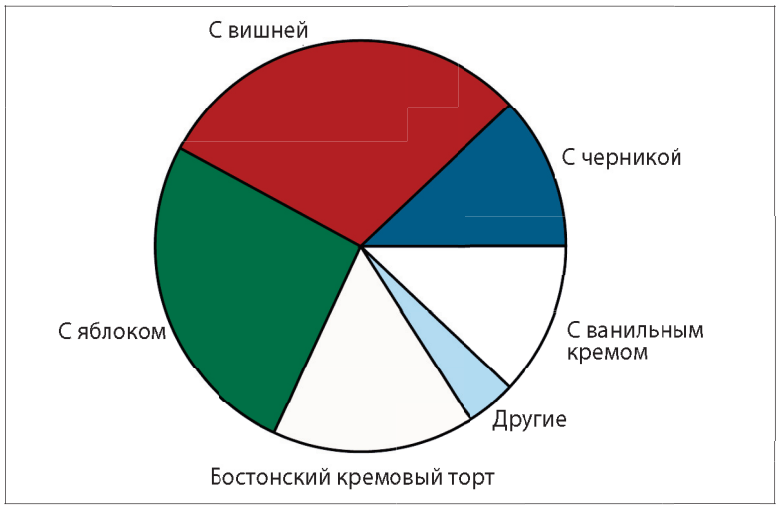


Рис. 5.8. Круговая диаграмма

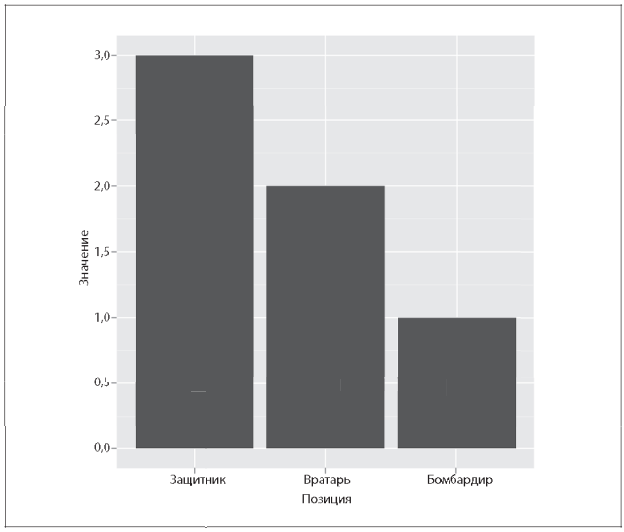


Рис. 5.9. Столбиковая диаграмма

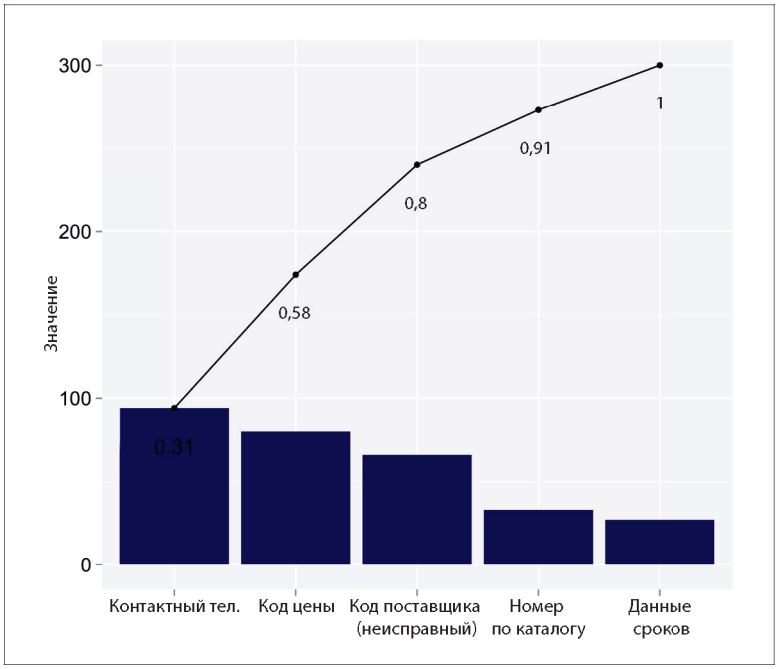


Рис. 5.10. Диаграмма Парето

Для визуализации двух переменных можно воспользоваться разными типами графиков.

	Непрерывная или дискретная переменная	**Категориальная переменная**
Категориальная переменная	Коробчатая диаграмма (box plot) Диаграмма с областями (area chart) Гистограмма распределения (range chart) Таблица (table chart)	Лепестковая диаграмма (spider/radar chart) Составная столбиковая диаграмма (stacked bar chart) Воронкообразный график (funnel chart)
Непрерывная или дискретная переменная	Диаграмма рассеяния (scatter plot) Линейный график (line graph) Карты и диаграмма Вороного (maps & Voronoi diagram) График плотности (density plot) Контурная диаграмма (contour plot)	Такие же, как в левом верхнем углу

(См. также рис. 7.5.)

Есть целый набор графиков для одновременного изучения трех переменных. Некоторые из них более общие и привычные (график поверхности (surface), пузырьковая диаграмма (bubble plots), 3D-диаграмма рассеивания (3D scatter)), а некоторые применяются для особых целей (см. the D3 gallery[1]).

В случае, когда одна из переменных — время (например, годы) или категориальная переменная, также можно использовать подход небольших множеств (small multiples), при котором создается решетка из одномерных или двумерных графиков (рис. 5.11).

Не ограничивайтесь использованием одного или двух типов диаграмм. Каждый из этих типов диаграмм выполняет свою задачу. Изучите их преимущества и недостатки и применяйте те из них, которые лучше всего отражают интересные сигналы, тренды или образцы. (Мы еще вернемся к некоторым из этих аспектов в главе 7.)

[1] URL: https://github.com/d3/d3/wiki/Gallery.

Там, где возможно, пользуйтесь командами, например `pairs()`, при автоматическом создании графиков и диаграмм для различных комбинаций переменных, которые вы можете быстро просмотреть в поисках интересных деталей или странностей, заслуживающих дополнительного внимания.

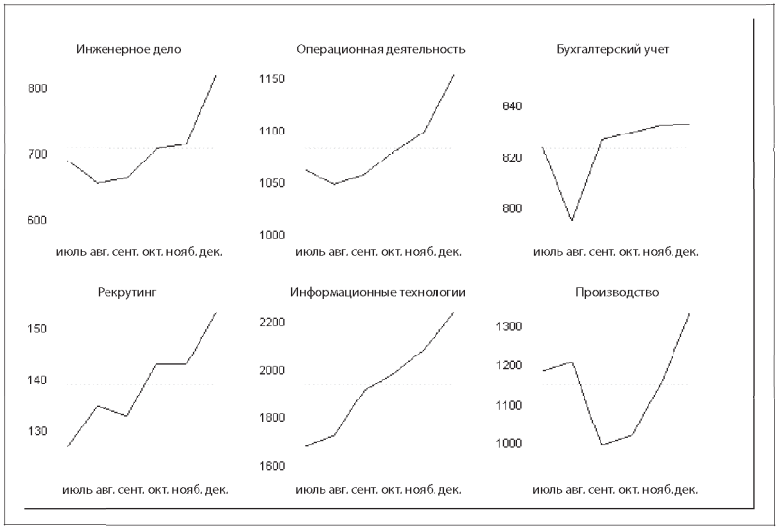


Рис. 5.11. Пример маленьких множеств

Источник: https://en.wikipedia.org/wiki/Small_multiple

ИНДУКТИВНЫЙ АНАЛИЗ

Описательный и разведочный виды анализа выступают под широкой зонтичной структурой описательной статистики: они *описывают* характеристики предлагаемого набора данных. Далее мы перейдем к другому основному направлению — статистическим исследованиям. Их цель заключается в *логическом извлечении* информации (параметры, распределение или взаимосвязи) о более широкой генеральной совокупности, из которой был взят набор данных. Кроме того, они обеспечивают основу для тестирования гипотез, на основе которых можно разрабатывать и проводить эксперименты для анализа нашего понимания внутренних механизмов и процессов.

Поскольку наша книга не учебник по статистике, в этом разделе мы лишь поверхностно проведем обзор вопросов, которые могут возникнуть, типов практических выводов, которые можно сформулировать, а также дополнительной ценности, которую можно получить благодаря

применению индуктивного анализа. Если вам требуется более подробная вводная информация по теме, настоятельно рекомендую ознакомиться с бесплатным ресурсом OpenIntro Statistics[1].

Зачем нужны статистические выводы? Как правило, мы делаем выводы обо всей генеральной совокупности на основе взятой из нее выборки, так как полный сбор данных бывает слишком дорогим, непрактичным, а иногда и просто невозможным. Возьмем, например, опрос граждан на выходе с избирательных участков, так называемый экзитпол. Невозможно опросить 125 млн избирателей, но вместо этого можно постараться получить качественную репрезентативную выборку и сделать точное умозаключение, каким мог быть результат, если бы были опрошены все избиратели. Также если вы обеспечиваете проверку качества производимой продукции и проводите испытания с *разрушением* опытного образца, очевидно, что вы не сможете протестировать подобным образом абсолютно всю продукцию, иначе вам просто нечего будет продавать.

Еще одна причина применения индуктивного анализа заключается в обеспечении объективности оценки расхождений и результатов. Предположим, вы решили провести кампанию для поощрения лояльности своих клиентов[2] и выбрали тысячу клиентов на основе общего критерия: например, каждый из них совершил не менее двух покупок за прошедший год и участвует в программе лояльности. Половине из отобранных клиентов (тестовая группа) вы отослали небольшой подарок с сообщением: «Просто потому, что мы любим своих клиентов, мы хотим преподнести вам этот скромный подарок». Вторая половина из отобранных клиентов (контрольная группа) не получила ничего. В течение следующих трех месяцев вы оцениваете число совершённых покупок, и описательный анализ показывает, что участники тестовой группы ежемесячно тратят на покупки в среднем на 3,36 долл. больше, чем участники контрольной группы. Что это означает? Очевидно, что это хорошо, но насколько надежны эти цифры? Получили бы мы похожий результат при повторном проведении эксперимента, или это просто случайность? Может быть, все объясняется тем, что один покупатель сделал крупный заказ? Статистические выводы позволяют оценить вероятность того, что это повышение покупательского спроса было просто случайностью, если при этом не наблюдалось реальных изменений внутренних образцов покупательского поведения.

---

[1] URL: https://www.openintro.org/stat/textbook.php.

[2] URL: http://brainsonfire.com/2013/02/12/7-awesome-examples-of-surprise-and-delight-that-will-blow-your-mind/.

Представьте, что вы отчитываетесь о результатах перед руководителем. На основе описательного анализа вы можете только констатировать результат: «Мы обнаружили разницу в объеме 3,36 долл./месяц, вектор движения правильный, и, кажется, это результаты кампании по поощрению лояльности клиентов». Однако на основе индуктивного анализа ваши выводы могут быть более убедительными: «Мы обнаружили разницу в объеме 3,36 долл./месяц, и вероятность того, что мы получили бы подобный результат без реального изменения в поведении покупателей, составляет всего 2,3%. Данные убедительно свидетельствуют, что это эффект от проведения кампании по поощрению лояльности клиентов». Или наоборот: «Мы обнаружили разницу, но при этом вероятность того, что этот результат случаен, составляет 27%. Вероятнее всего, кампания не была эффективной, по крайней мере, для данного конкретного показателя». Как с позиции аналитика, так и с позиции руководителя можно утверждать, что индуктивный анализ имеет бóльшую ценность и оказывает более значительное влияние на деятельность компании.

Статистические выводы обеспечивают ответы на приведенные ниже типы вопросов (но не ограничиваются ими).

Стандартная ошибка, доверительный интервал, статистическая погрешность

Насколько можно быть уверенным в этом среднем выборочном или в доле выборки? Насколько будет отличаться значение, если провести эксперимент повторно?

Математическое ожидание по одной выборке

Насколько полученное среднее выборочное отличается от ожидаемого значения?

Разница средних значений по двум выборкам

Насколько сильно отличаются средние значения по двум выборкам? (Говоря более техническим языком, какова вероятность, что мы бы наблюдали эту разницу средних значений или выше, будь верна нулевая гипотеза про отсутствие разницы между средними значениями по генеральной совокупности по двум выборкам?)

Вычисление размера выборки и анализ статистической мощности

Каким должен быть минимальный размер выборки, учитывая, что мне уже известно о процессе, чтобы достигнуть определенного

уровня уверенности в качестве данных? Эти типы статистических инструментов важны для планирования A/B-тестирования (подробнее об этом в главе 8).

Распределение данных

Соответствует ли распределение значений в этой выборке нормальному (конусообразному) распределению? Вероятно ли, что у этих двух выборок будет одинаковое исходное распределение?

Регрессия

Предположим, я провел тщательно разработанный эксперимент, в котором системно изменял одну (независимую) переменную, контролируя при этом максимально возможное число других факторов, после чего я построил прямую регрессии. Насколько я могу быть уверен в этой прямой? Насколько высока вероятность ее изменения (угол наклона и точка пересечения) при многократном повторении эксперимента?

Критерий соответствия и ассоциированности

В случае с категориальной переменной (например, категория продукта), соответствует ли частота или число (например, покупок) ожидаемой относительной частоте? Наблюдается ли взаимосвязь между двумя переменными, одна из которых категориальная?

Несмотря на краткость приведенного обзора, надеюсь, вы смогли разглядеть потенциальную ценность того набора инструментов, с помощью которого делаются статистические выводы. Он позволяет разрабатывать эксперименты и получать более объективный анализ данных, снижая количество ложноположительных результатов, происходящих из-за чистой случайности.

ПРОГНОСТИЧЕСКИЙ АНАЛИЗ

Делать прогнозы чрезвычайно сложно,
особенно относительно будущего.
приписывается Нильсу Бору

Прогностический анализ строится на индуктивном анализе. Цель в том, чтобы изучить взаимосвязи между переменными на основе существующего набора данных и разработать статистическую модель,

способную прогнозировать значения для новых, неполных или будущих точек данных.

На первый взгляд это кажется магией вуду, не меньше. В конце концов, мы не имеем ни малейшего представления, когда следующее мощное землетрясение разрушит Сан-Франциско (сроки имеющегося предсказания уже прошли), где и когда в следующем сезоне образуются ураганы или сколько будут стоить акции Apple в понедельник утром (если бы я мог сделать такой прогноз, то не писал бы сейчас эту книгу). Реальность такова, что мы не в состоянии точно предсказать какие-то неожиданные события и катастрофы, так называемых черных лебедей[1]. При этом во многих аспектах бизнеса и других областях знаний есть достаточные сигналы, с обработкой которых прогностический анализ отлично справляется. Например, в 2008 году Нейту Сильверу удалось предсказать результаты выборов в Сенат и победителей в 49 штатах из 50.

В сфере розничной торговли могут наблюдаться устойчивые закономерности. На рис. 5.12 приводится четкая и предсказуемая кривая (синяя сверху) ежегодных продаж солнечных очков, которая достигает пика в июне-июле и находится на спаде в ноябре и январе (предположительно небольшой ее рост наблюдается в декабре во время сезонной распродажи). Похожая кривая, но со смещением на шесть месяцев, отражает ежегодные продажи перчаток: ее пик приходится на декабрь. Таким образом, на основе результатов прогностического анализа можно разработать планы, когда производить или покупать товары, какой объем товаров производить или покупать, когда организовать доставку в магазины и так далее.

Помимо временны́х рядов прогностический анализ также способен делать прогнозы, к какому *классу* может относиться объект анализа. Например, на основе информации о размере заработной платы, истории покупок, оплаченных кредитной картой, истории оплаты (или неоплаты) счетов того или иного человека можно вычислить степень кредитного риска. Или на основе записей в Twitter, содержащих краткую оценку фильма, каждый из которых был отмечен пользователем положительно («фильм понравился») или отрицательно («отвратительный фильм»), можно разработать модель, прогнозирующую эмоциональную окраску — положительную или отрицательную — новых записей,

[1] Taleb N. N. The Black Swan. The Impact of the Improbable (New York: Penguin Press, 2007). Издана на русском языке: Талеб Н. Черный лебедь. Под знаком непредсказуемости. М. : Азбука-Аттикус : КоЛибри, 2016. *Прим. ред.*

например, таких как «спецэффекты в фильме просто классные», которые не вносились в модель ранее.

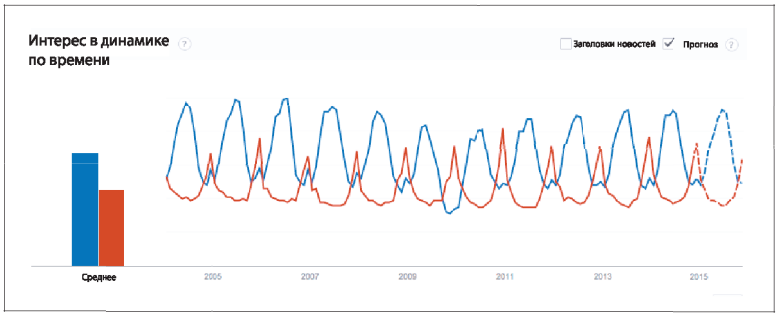


Рис. 5.12. Инструмент Google Trends отражает предсказуемую сезонную закономерность интереса к солнечным очкам (верхняя синяя кривая) и перчаткам (нижняя красная кривая) в период 2004–2014 годов и прогноз на год, до 2015-го

Существует множество приложений, использующих прогностическую аналитику, и они весьма заметны на рынке. Ниже приведено несколько примеров.

Прогнозы, формирующие основу сервиса как такового

Приложения для знакомств

Качественные приложения для поиска новых знакомых могут повысить степень удовлетворенности потребителей.

Приложения для игры на бирже (на риск пользователя!)

Отслеживая движение цен на акции и определяя закономерности, с помощью специальных алгоритмов можно попытаться покупать на спаде, продавать на пике и максимизировать рентабельность вложенных средств.

Прогнозы, обеспечивающие более высокий уровень обслуживания для клиентов

Спам-фильтры

Обнаружение и фильтрация спама («Купите “Виагру” онлайн») от не спама («Запланированная встреча с генеральным директором») делает работу с электронной почтой более эффективной, а пользователя — более счастливым.

Рекомендации по контенту

Качественные рекомендации, что можно посмотреть (Netflix), гарантируют возврат пользователей и снижают количество пользователей, отказавшихся от услуг.

Общение в социальных сетях

Сервис LinkedIn «Люди, которых вы можете знать» повышает эффективность пользования социальной сетью и обеспечивает более высокую ценность для пользователей и более ценные данные для социальной сети.

Прогнозы, способные обеспечить более высокий уровень конверсии и размер корзины

Кросс-продажи и увеличение объема покупки

Даже самые простые рекомендации, основанные на ассоциациях, например «Пользователи, которые купили DVD “Холодное сердце”, также покупают “Русалочку”» (Amazon), увеличивают объем продаж, а некоторым пользователям значительно облегчают и ускоряют процесс совершения покупок.

Рекламные объявления и купоны

Изучение истории покупок пользователя, а также прогнозирование его потенциальных интересов или намерений, может способствовать более релевантному отображению рекламных объявлений или более эффективному предложению купонов (например, от компании Tesco, далее мы поговорим об этом подробнее).

Прогнозы, способствующие улучшению стратегии

Одобрение от банка

Прогноз, у кого из заемщиков потенциально могут возникнуть трудности с выплатой взятых на себя обязательств, можно включить в процесс одобрения кредитных заявок, что снизит риск невозврата кредита.

Прогнозирование в работе органов правопорядка

Можно делать прогнозы относительно того, где могут вспыхнуть беспорядки, и принимать решения, куда и когда отправить полицейские наряды.

Прогнозирование активности пользователей

Благодаря прогнозированию наплыва или активности пользователей, например, что во время «Суперкубка» может произойти резкое увеличение количества сообщений в Twitter, можно заранее расширить технические мощности, чтобы предотвратить сбой в работе сервиса.

Политические кампании

Качественное прогнозирование намерений избирателей (голосовать / не голосовать, за демократов / за республиканцев / не определился) и ежедневное обновление данных привело к повышению эффективности в работе со СМИ, во взаимодействии с избирателями и в сборе пожертвований на проведение избирательной кампании, что в значительной мере обеспечило успех президентской кампании Барака Обамы.

Это всего лишь несколько примеров. Для получения более подробного обзора по теме прогностического анализа я рекомендую книгу Джона Сигела Predictive Analytics (John Wiley & Sons), в частности табл. 1–9.

Итак, как проводится прогностический анализ? Для этого существует целый ряд инструментов и подходов. Самая простая из возможных моделей — прогнозировать, что завтра будет таким же, как сегодня. Этот подход может сработать в случае медленно изменяющихся явлений, например, когда речь идет о погоде в Южной Калифорнии, но не в случае с волатильными системами, например такими, как цена на акции. Регрессия — самая обширная семья статистических инструментов. Для работы с разными характеристиками данных применяют разные виды регрессии (лассо-регрессию, гребневую, робастную и так далее). Особенный интерес представляет логистическая регрессия, которую можно применять для прогнозирования классов. Например, если раньше для определения категории спам / не спам использовалась модель наивного байесовского классификатора, то сегодня чаще применяется логистическая регрессия. К другим техникам и так называемому машинному обучению относятся нейронные сети, деревья решений и регрессии, алгоритм машинного обучения «Случайный лес», метод опорных векторов, метод *k* ближайших соседей.

Прогностический анализ весьма эффективен, но не обязательно сложен. Наиболее сложное в нем — получить качественный набор данных. При разработке классификатора часто это означает ручной контроль над данными, например маркировку набора сообщений в Twitter как положительных или отрицательных, что может быть особенно трудоемко.

Однако при наличии этих данных с хорошей библиотекой, такой как scikit-learn[1], для составления базовой модели потребуется буквально несколько строк кода. При этом для получения *хорошей* модели часто требуется приложить больше усилий, провести больше итераций, а также процесс генерирования признаков (feature engineering). Признаки — вводные данные для модели. Они могут включать основные собранные данные, например количество заказов, простые производные переменные, такие как «Заказ был сделан в выходные? Да/нет», а также более сложные абстрактные признаки, такие как «коэффициент похожести» двух фильмов. Генерация признаков — это и искусство, и наука, и она зависит от степени владения профессиональными знаниями.

Наконец, для проведения прогностического анализа не требуется большого объема данных. Объем базы данных, на основе которой Нейт Сильвер составлял прогнозы по итогам предвыборной кампании 2008 года, был всего 188 тыс. единиц (см. презентацию Оливера Гризела, в которой подтверждаются эти цифры и приводится хороший краткий обзор прогностического анализа[2]). Основную роль сыграло то, что Сильвер располагал множеством самых разных источников и данных опросов, каждый из которых в чем-то был ошибочным и необъективным, тем не менее в совокупности они относительно точно отразили действительность. Подтверждено на практике, по крайней мере для определенных классов проблем, что большой объем данных позволяет обходиться простыми моделями[3] (см. приложение А).

Резюмируя сказанное, прогностический анализ — мощный инструмент в арсенале компании с управлением на основе данных.

КАУЗАЛЬНЫЙ (ПРИЧИННО-СЛЕДСТВЕННЫЙ) АНАЛИЗ

Вероятно, каждый из нас знает утверждение: «Корреляция не подразумевает причинно-следственных отношений»[4]. Если вы проведете сбор данных, а затем разведочный анализ, чтобы выявить интересные взаимосвязи между переменными, то, скорее всего, что-нибудь обнаружите. Однако даже если между двумя переменными наблюдается очень сущес-

[1] URL: http://scikit-learn.org/stable/.

[2] URL: https://speakerdeck.com/ogrisel/predictive-analytics.

[3] Fortuny E. J. de, Martens D. and Provost F. Predictive Modeling with Big Data: Is Bigger Really Better? Big Data 1, no. 4 (2013): 215–226. URL: http://online.liebertpub.com/doi/full/10.1089/big.2013.0037.

[4] Если не верите, проверьте ложные корреляции (например, объем потребления сыра в США коррелирует с количеством людей, умерших от того, что запутались в собственном постельном белье). URL: http://www.tylervigen.com/spurious-correlations.

твенная корреляция, это не означает, что одна из них обусловливает другую. (Например, уровень холестерина-ЛПВП обратно пропорционален вероятности развития сердечно-сосудистых заболеваний: чем выше уровень этого «хорошего» холестерина, тем лучше. При этом препараты, повышающие уровень холестерина-ЛПВП, никак не влияют на предотвращение сердечно-сосудистых заболеваний. Почему? Потому что холестерин-ЛПВП представляет собой побочный продукт нормальной сердечной деятельности, а не ее причину.) Таким образом, у подобного апостериорного анализа есть серьезные ограничения. Если вы действительно хотите понять систему и точно узнать, какими рычагами влияния на фокусные переменные и показателями вы обладаете, тогда вам требуется разработать причинно-следственную модель.

Основная идея похожа на ту, что была в описанном ранее примере с поощрением лояльности клиентов: провести один или серию экспериментов с изменением одного параметра и контролем максимального количества всех остальных. Например, можно провести эксперимент с электронной рассылкой клиентам, в которой вы протестируете тему сообщения. При прочих равных условиях (то же самое содержание, время отправки и так далее) с единственной разницей в теме, если вы отметите, что уровень просмотра сообщения с другой темой гораздо выше, у вас есть все основания сделать вывод, что именно тема сообщения — причина интереса к нему.

У этого эксперимента есть свои ограничения, так как, несмотря на то что он подтверждает влияние фактора темы сообщения, неясно, какое именно слово или фраза вызвали отклик пользователей. Чтобы это выяснить, требуется проведение дополнительных экспериментов. Рассмотрим более количественный пример: время отправки сообщения может оказать серьезное влияние на уровень просмотра. Чтобы это проверить, можно провести контролируемый эксперимент с вариантами (сделать отправку электронной рассылки по частям в 8, 9, 10 часов утра и так далее) и проанализировать, как время отправки сообщения повлияло на уровень просмотра. Так вы сможете прогнозировать (интерполировать) предполагаемый уровень просмотра сообщения, отправленного в 8:30 утра.

ЧТО ВЫ МОЖЕТЕ СДЕЛАТЬ?

Рекомендация аналитикам. Вам стоит стремиться действовать в двух направлениях — «точить топор» и расширять арсенал инструментов. Вы станете более эффективным и ценным специалистом, кроме того, это будет инвестицией в себя и в развитие вашей карьеры. Оцените статистические навыки и навыки визуализации данных,

которыми вы сейчас пользуетесь. Как вы можете их улучшить? Например, если вы освоите среду R, поможет ли это вам быстрее и эффективнее проводить разведочный анализ? Окажет ли более глубокий аналитический подход более важное влияние на ваш проект? Что вам необходимо, чтобы овладеть новым навыком?

Рекомендация руководителям. Обращайте особое внимание на ситуации, в которых применение дополнительных видов аналитической работы способно обеспечить более глубокие выводы и повлиять на эффективность деятельности компании. Если отсутствие товара на складе становится проблемным местом цепочки поставок, можно ли исправить эту ситуацию с помощью прогнозных моделей? Можно ли проводить больше экспериментов, которые углубят институциональные знания причинных факторов? Стимулируйте специалистов по работе с данными, чтобы они повышали квалификацию, и всячески их в этом поддерживайте. Позвольте им опробовать новые программные средства, которые могут облегчить их работу и сделать ее более эффективной.

Подобные эксперименты обеспечивают более глубокое понимание системы и причинно-следственных взаимосвязей, что можно использовать при составлении прогнозов и планировании кампаний и других изменений, цель которых — улучшить и без того хорошие показатели, которых кто-то только стремится достичь. На их основе также можно строить имитационные модели, которые можно применять для оптимизации системы. Например, можно смоделировать цепочку поставок и изучить, как разные варианты схемы и условий пополнения склада влияют на дефицит товаров на складе или на совокупные расходы на транспортировку и хранение товаров. Этот вид деятельности отражен в правом верхнем углу матрицы Дэвенпорта в табл. 1.2. Это наивысший уровень аналитики. Принимая во внимание контролируемый, научный характер сбора данных на протяжении определенного периода, а также высокую эффективность подобных каузальных моделей, они становятся, по словам Джеффри Лика, «золотым стандартом» анализа данных.

С точки зрения ведения бизнеса вся эта бурная деятельность по анализу данных и разработке моделей проводится не ради самой деятельности и не по прихоти высшего руководства. Ее цель — поддержка основных показателей, таких как уровни просмотров, конверсии, наконец, показатель выручки. Поэтому критически важно, чтобы эти основные показатели были правильными и были качественно разработаны. В противном случае вы будете оптимизировать не то, что надо. Учитывая важность качественной разработки показателей, подробнее остановимся на этом вопросе в следующей главе.

ГЛАВА 6

Разработка показателей

Когда не знаешь, куда идешь, то, скорее всего, окажешься где-нибудь еще.
Йоги Берра

Считайте, что поддается подсчету, измеряйте, что поддается измерениям, а неизмеряемое делайте измеряемым.
Галилео Галилей

В компании с управлением на основе данных должна быть четкая стратегия, то есть направление развития бизнеса, а также конкретный набор основных показателей — ключевых показателей эффективности деятельности (KPI) — для отслеживания, в верном ли направлении и насколько успешно идет развитие бизнеса. Ответственность за достижение этих KPI ложится на бизнес-единицы или подразделения, где могут быть определены дополнительные KPI специально для этого подразделения. Это завершает набор операционных и диагностических показателей, на основе которых контролируется выполнение задач, программ, тестов и проектов, ведущих к выполнению KPI.

Учитывая сказанное, чрезвычайно важна качественная разработка показателей. Они выполняют такую же роль, как точный компас. Вряд ли вы захотите следовать *стратегическому* показателю, указывающему, что вы продвигаетесь в желаемом юго-восточном направлении, когда на самом деле вы идете на северо-восток, или *операционному* показателю, отражающему ежегодный рост конверсии на 5%, когда на самом деле никакого роста нет. Точно так же вы не захотите руководствоваться неверным *диагностическим* показателем, который не в состоянии как можно раньше проинформировать вас о том, что ваш сайт на грани краха. Показатели, кроме того, представляют собой результаты экспериментов и A/B тестов, которые при правильном подходе вносят весомый вклад в каузальный анализ, что, как мы обсуждали в предыдущей главе,

может стать отличной основой для формулирования выводов и стратегий на основе данных. Эту идею удачно сформулировал Дэвид Скок:

> Один из способов оценить работу компании — представить ее в виде автомата, выдающего определенный объем продукции, с рычагами, с помощью которых управленческая команда способна контролировать его работу. У слабой команды ограниченное понимание, как работает ее автомат и какие у нее есть рычаги влияния. Чем лучше управленческая команда, тем лучше она понимает схему работы автомата и то, как можно оптимизировать его работу (на какие рычаги нажать). При разработке показателей мы стремимся улучшить свое понимание автомата и схемы его работы. Качественно разработанные показатели будут способствовать повышению результативности работы на выходе[1].

В этой главе мы поговорим о разработке показателей. Начнем с общих вопросов, а затем перейдем к KPI. Однако мы лишь поверхностно обсудим вопрос выбора показателей, так как полноценная дискуссия выходит за рамки этой книги. Кроме того, этому важному этапу посвящен целый ряд убедительных концепций, таких как сбалансированная система показателей, всеобщее управление качеством (TQM), призма эффективности и концепция Tableau de Bord («Бортовое табло»).

Разработка показателей

При выборе или разработке показателей следует руководствоваться несколькими принципами. В идеальном мире показателям должны быть присущи несколько характеристик.

ПРОСТОТА

Разрабатывайте показатель, чтобы он был «таким простым, как только возможно, но не проще» (Эйнштейн).

Какое из этих определений будет понятнее вашим коллегам?

Клиент — человек, который отдает деньги и получает один из товаров компании.

[1] URL: http://www.forentrepreneurs.com/designing-startup-metrics-to-drive-successful-behavior/.

Клиент — человек, купивший товар,

— за исключением покупки подарочного сертификата;

— за исключением тех, кто вернул товар в течение 45 дней с момента покупки с полным возвратом стоимости;

— включая тех, кто активирует подарочный сертификат.

Надеюсь, вы уловили основную мысль.

Простые показатели, по определению, просто объяснить, это означает следующее:

— их суть проще донести до других людей: возникает меньше непонимания;

— их проще реализовать: выше вероятность, что их рассчитают правильно;

— они с большей вероятностью поддаются сравнению с показателями других подразделений или компаний.

Конечно, есть множество обоснованных причин, почему требуется добавить дополнительный бизнес-критерий или пограничный случай для создания более сложного показателя. Возможно, вам необходимо фильтровать источники, чтобы они не содержали необъективные или резко отличающиеся данные. Или вам может понадобиться показатель, по которому выделяется конкретная подгруппа данных, например те случаи обслуживания клиентов, которые стоили компании дороже всего.

Каждый случай следует рассматривать по существу, но постарайтесь избегать дополнительных сложностей с редкими пограничными случаями, которые не добавляют особой ценности для бизнеса и лучшего понимания этого показателя.

Вывод: не стоит чрезмерно усложнять показатели.

ЕДИНЫЙ СТАНДАРТ

По возможности руководствуйтесь общепринятыми стандартами. Например, имея единый, четко определенный показатель отказов, используйте его в своей деятельности, если только у вас нет веской причины для создания своего собственного варианта этого показателя. Если в розничной торговле проходимость торговой точки считается по количеству вышедших из магазина, используйте этот показатель,

а не считайте количество вошедших, даже если эти показатели сопоставимы концептуально и по своим значениям. Например, при отслеживании ежемесячной активности пользователей Facebook включает в подсчет только тех, кто залогинился на сайте, в то время как Yelp включает и эту категорию и тех, кто использует гостевой доступ.

Применение общепринятых стандартов вызовет меньше непонимания, особенно у коллег, пришедших из других компаний. К тому же вам будет легче сравнивать свои показатели с показателями других компаний отрасли, то есть анализировать результаты своей работы относительно наиболее эффективных практик в отрасли.

Еще важнее, чтобы все показатели были стандартизированы в рамках одной компании. Мне доводилось наблюдать, как разные подразделения были уверены, что применяют один и тот же показатель, и даже описывали его в одинаковых терминах, но на практике реализация этого показателя в таблицах или системах этих подразделений значительно различалась. Их цифры не совпадали, что приводило к ожесточенным спорам.

Оптимальный вариант — иметь единый централизованный, автоматический, документально подтвержденный «источник истины», из которого бы черпали информацию разные подразделения. Тогда вы сможете использовать результаты анализа и выводы коллег в полной уверенности, что вы сравниваете подобное с подобным. В этом случае становится проще создать единое хранилище результатов аналитической работы и корпоративных знаний о причинных факторах в бизнесе (или о рынке), которому можно доверять и использовать.

Вывод: применяйте общепринятые показатели, если только у вас нет веских причин от них отклониться. При использовании нестандартных показателей зафиксируйте документально, как и почему они нестандартные.

ДОСТОВЕРНОСТЬ

Показатели должны быть достоверными. Это означает, что их среднее числовое значение должно быть приближено к истинному теоретическому среднему значению (см. рис. 6.1). Если использовать метафору стрельбы из лука, то стрела должна попасть точно в мишень.

Возьмем, например, объем выручки от продаж на Amazon. Показатель среднего объема выручки *за исключением суммы от продажи книг* — неточное среднее значение совокупного объема выручки от всех продаж.

Этот показатель необъективен. В главе 2 мы уже обсуждали примеры, когда отсутствующие данные приводили к искажению общей картины. Например, средний уровень удовлетворенности клиентов не отражает действительность, если недовольные клиенты из-за задержки доставки товара пропустили дедлайн по опросу и не предоставили свои ответы. В этом примере показатель степени удовлетворенности клиентов завышен по сравнению с его истинным более низким значением.

При разработке показателей постарайтесь учесть все потенциальные источники искажения, как в данных, так и в самом показателе. В главе 2 мы обсуждали некоторые источники необъективности при сборе данных. С точки зрения показателя подумайте обо всех возможных фильтрах при сборе данных, а также о любых скрытых или устаревших «поправочных коэффициентах».

Представьте себе стрелка, который готовится стрелять по дальней мишени и пользуется оптическим прицелом. При стрельбе следует учесть силу и направление ветра, влияющие на траекторию движения пули. Поэтому стрелок регулирует прицел — «поправочный коэффициент» — с поправкой на ветер. При этом если сила или направление ветра изменятся, то эта поправка окажется устаревшей, пули больше не попадут в цель. Внешние обстоятельства часто меняются, а потому необходимо внимательно следить за актуальностью действующих моделей и поправочных коэффициентов.

То же самое верно и в бизнесе. В Warby Parker мы используем электронные устройства для подсчета количества посетителей, вошедших и вышедших из наших розничных магазинов. Одно из возможных применений этих данных — для вычисления показателя конверсии торговой точки, то есть количества посетителей, зашедших в магазин и совершивших какую-нибудь покупку. В одном из таких магазинов персонал может попасть на склад с товаром и вернуться в торговый зал только через главный вход: эти передвижения точно так же считались электронными приборами, из-за чего показатель конверсии получался заниженным. Мы постарались исправить ситуацию, разработав статистическую модель, которая для конкретного дня недели и конкретного уровня занятости оценивала соотношение трафика персонала и посетителей магазина в качестве корректирующего фактора. В результате показатель конверсии стал гораздо более реалистичным. Следует учесть, что подобные модели могут терять свою актуальность при изменении внешних условий, например покупатели могут быть более мотивированы совершать покупки по выходным. Нужно либо

периодически перенастраивать модель, либо, как мы пробуем делать сейчас, использовать более совершенные технологии, способные отличить персонал от посетителей и не включать сотрудников при подсчете трафика.

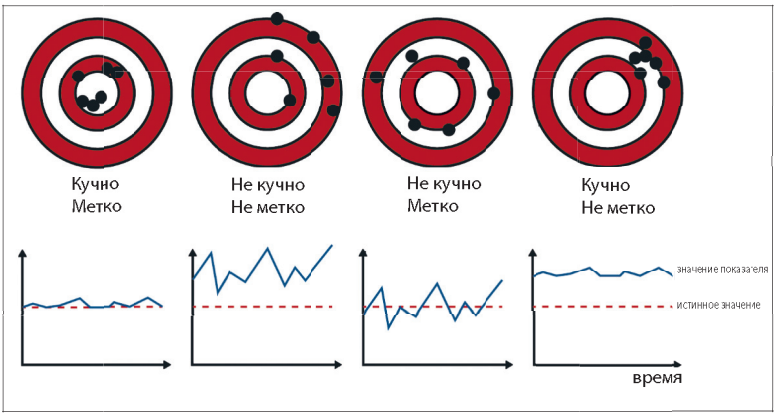


Рис. 6.1. Точность (в стрельбе есть такой термин, как «кучность» — группировка точек падения снарядов на ограниченной площади) и достоверность (по аналогии со стрельбой это меткость попадания в мишень) на примере двухмерных данных. Недостоверный показатель необъективен, так как его среднее значение системно отличается от истинного среднего значения. Точность показателя отражает его вариативность: насколько будет отличаться среднее значение, если вы повторите эксперимент несколько раз и соберете новые выборки такого же размера

ТОЧНОСТЬ

Показатели должны отличаться точностью. Это означает, что при повторении эксперимента в тех же самых условиях значения должны получаться такими же. По аналогии со стрельбой это можно назвать кучностью: все попадания в мишень должны быть рядом на ограниченной площади. Один из инструментов, или рычагов, для контроля точности — размер выборки. Чем больше выборка, тем меньше стандартная ошибка. Однако эта взаимосвязь не линейная. Так как стандартная ошибка среднего значения равна стандартному отклонению, деленному на квадратный корень размера выборки, чтобы уменьшить стандартную ошибку в два раза, нужно в *четыре раза* увеличить размер выборки.

Сочетание достоверности (меткости попадания в мишень) и точности (кучности стрельбы) показано на рис. 6.1. Если у вас нет подтвержденной справочной информации, вы можете не понять, что ваши показатели недостоверны. Однако вы, скорее всего, рано или поздно поймете, если ваши показатели не отличаются точностью (нестабильны).

Вывод: стремитесь к достоверности и точности показателей и учитывайте издержки и преимущества крупных выборок.

ОТНОСИТЕЛЬНЫЕ ИЛИ АБСОЛЮТНЫЕ ПОКАЗАТЕЛИ

Очень важное решение — относительные или абсолютные показатели следует применять. Этот выбор определяет разработку показателей, которые при одном сценарии показывают очень разные картины.

Представьте, что в какой-то компании ведется классификация клиентов и 25% от общего количества относятся к категории VIP (например, они приобрели продукцию компании на сумму больше 1 тыс. долл.). Через полгода у этой компании только 17% VIP-клиентов. Черт, что случилось? Они что, ушли? Как все исправить?

Предположим, что в этот период усилия компании были сосредоточены на привлечении новых клиентов. Тогда, вероятно, общее количество клиентов увеличилось (показано оранжевым на рис. 6.2), а количество VIP-клиентов могло остаться тем же, при этом их пропорция уменьшилась. Фактически вполне возможно даже, что количество VIP-клиентов тоже увеличилось, но при этом пропорция стала ниже.

И наоборот, предположим, что через полгода мы наблюдаем значительное увеличение количества VIP-клиентов и их пропорции. Это может отражать здоровый рост клиентской базы, но, с другой стороны, роста клиентской базы может и не быть, если усилия компании были сосредоточены исключительно на возвращении покупателей и увеличении количества повторных покупок (рис. 6.2, внизу). (Для многих компаний второй сценарий с увеличением количества повторных покупок более предпочтителен по сравнению с увеличением клиентской базы, так как стоимость привлечения новых клиентов, как правило, слишком высока.)

Как видите, выбор между применением абсолютных (количество VIP-клиентов) или относительных (их пропорция) показателей может привести к очень разным интерпретациям.

Вывод: тщательно взвесьте, что вы хотите узнать, и выберите абсолютный или относительный показатель, который будет адекватно отображать нужные вам изменения.

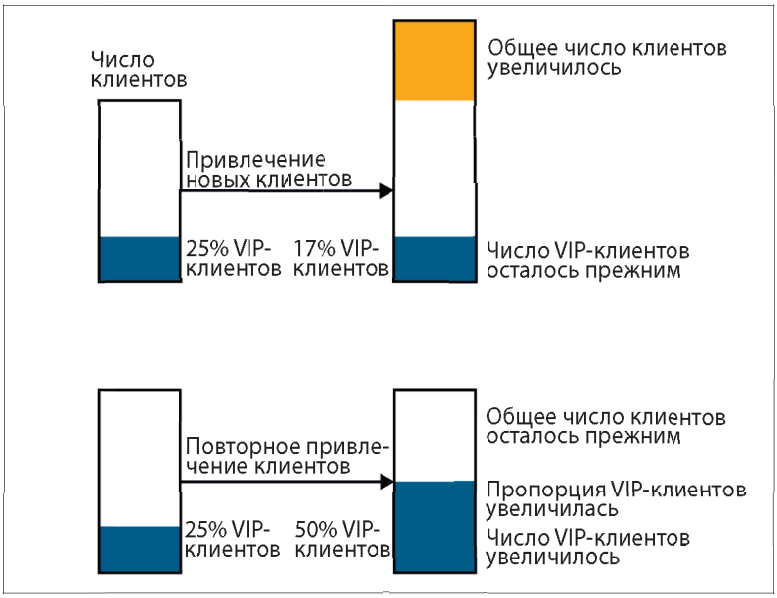


Рис. 6.2. У компании 25% VIP-клиентов. В верхнем сценарии компания сосредоточила усилия на привлечении новых клиентов (показано оранжевым). Это привело к увеличению общего количества клиентов, количество VIP-клиентов осталось прежним, но пропорция уменьшилась. В нижнем сценарии компания сосредоточила усилия на работе с текущими клиентами. Пропорция и количество VIP-клиентов стали выше, но общего увеличения клиентской базы не произошло

РОБАСТНОСТЬ[1]

Определяйте статистически робастные показатели, то есть те, что относительно нечувствительны к отдельным резко отличающимся значениям.

Рассмотрим следующий пример из San Francisco Chronicle:

> Средняя заработная плата специалистов технического профиля в центральной части полуострова Сан-Франциско (округ Сан-Матео) в прошлом году составила 291 497 долл. Возможное объяснение отклонения: глава компании Facebook Марк Цукерберг получил всего один доллар в качестве

[1] Робастность (от *англ.* robust — «крепкий», «твердый», «устойчивый») — свойство статистического метода, характеризующее независимость влияния на результат исследования различного рода выбросов, устойчивость к помехам. *Прим. перев.*

> зарплаты, но заработал 3,3 млрд долл. на опционах на покупку акций в 2013 году. Если вычесть 3,3 млрд долл. из общей суммы, то среднее значение получится примерно 210 тыс. долл.[1]

Использовать среднее значение в данном случае не следует, учитывая высокую степень позитивной асимметрии в данных по заработной плате. Среднее значение получается существенно завышенным (более чем на 35%) из-за одной резко отличающейся переменной. В данном случае гораздо рациональнее выбрать показатель медианы, так как он более устойчив к резко отличающимся значениям и лучше отражает средние данные.

Стоит отметить, что в некоторых случаях могут понадобиться показатели, которые особенно чувствительны к пограничным значениям. Пиковая нагрузка на веб-сайт должна охватывать редкие максимальные значения, которые должны быть включены в диапазон. Оценить или визуализировать робастность можно с помощью повторной выборки. Возьмите набор данных и вычислите показатель. Повторите расчеты несколько раз, заменяя набор данных; получив ряд значений показателя, составьте их распределение. Насколько это распределение отличается от того, что вы ожидали или хотели бы увидеть?

Вывод: примените разведочный анализ (например, постройте гистограмму или диаграмму рассеяния), чтобы лучше понять данные, и на его основании выберите робастные показатели.

ПРЯМАЯ СВЯЗЬ

Постарайтесь выбирать показатели, которые непосредственно измеряют интересующий вас процесс. К сожалению, не все можно измерить и оценить количественно, поэтому иногда приходится довольствоваться косвенными или приближенными показателями.

Кэти О'Нейл привела наглядный пример, как результаты тестов учеников приблизительно отражают качество обучения[2]. Чем больше расстояние между самим процессом и приближенным показателем, тем менее достоверным и полезным будет его значение. В результате вы можете начать оптимизировать приближенный показатель, что может оказаться совсем не тем, что вы действительно хотите оптимизировать.

---

[1] URL: http://blog.sfgate.com/pender/2014/08/21/these-tech-worker-wages-will-astound-you/.

[2] URL: http://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf.

Сьюзан Веббер рассказала о тестировании вкусов кока-колы и о выпуске на рынок нью-кок в 1980 году[1]. Компания провела маркетинговые исследования, которые показали в высшей степени положительные результаты, даже по сравнению с традиционной кока-колой. Однако когда новый продукт вывели на рынок, его продажи провалились. Почему?

Покупатели сочли напиток слишком сладким. Дело в том, что при тестировании вкуса в ходе маркетинговых исследований участники фокус-группы пробовали напиток маленькими глотками, в результате чего степень его сладости не так раздражала. Если бы они пробовали напиток «как в жизни» (сделали бы большой глоток жарким днем), то оптимизировали бы свое восприятие в соответствии с действительностью.

Вывод: везде, где возможно, оснащайте свои процессы и системы контрольно-измерительными средствами и старайтесь максимально избегать приближенных показателей. Не всегда стоит идти по пути наименьшего сопротивления и использовать данные, оказавшиеся под рукой. Сконцентрируйтесь на данных, которые вам следовало бы собрать и использовать, если они в большей степени отвечают вашим потребностям.

Ключевые показатели эффективности

Ключевые показатели эффективности (KPI) представляют собой набор значений самого высокого уровня, связанных со стратегическими целями компании. Они помогают определить и отследить направление, в котором развивается бизнес, и позволяют достигать намеченных целей. Как уже было сказано, эти показатели обеспечивают кораблю движение верным курсом.

Авинаш Кошик, ведущий мировой эксперт в области веб-аналитики, называет KPI «показателями, которые помогают понять, насколько эффективно вы действуете относительно своих целей»[2].

Он подчеркивает два краеугольных камня этого определения — показатели и цели, — так как KPI связывают их воедино. Примеры KPI: «Повысить узнаваемость бренда на 10%», «Удвоить количество активных пользователей к концу года», «Увеличить онлайн-конверсию на 5% во втором квартале».

---

[1] URL: http://www.auroraadvisors.com/articles/Webber-Metrics.pdf.

[2] URL: https://www.kaushik.net/avinash/rules-choosing-web-analytics-key-performance-indicators/.

Для KPI критически важны перечисленные ниже аспекты.

KPI должны быть четко определены

Не должно быть никакой двусмысленности в понимании основных показателей, к которым стремится компания. Показатель следует четко определить, у него должно быть конкретное целевое значение и обозначенный или стандартный срок (обычно конец года).

KPI должны быть измеряемыми

Ключевые показатели эффективности должны иметь числовое значение. Вам необходима возможность измерить прогресс в количественном выражении за определенный период времени. Иными словами, это должна быть иголка, которую можно передвигать с места на место, а не двоичное значение. Главный специалист США по анализу данных (US Chief Data Scientist) Ди Джей Патиль в своей книге Building Data Science Teams[1] отметил: «Как оказалось, все компании, в которых на высшем уровне развито управление на основе данных, придерживаются одного правила: если что-то нельзя измерить, это невозможно исправить».

KPI должны содержать цели

«Повысить выручку» — это плохо сформулированный ключевой показатель эффективности, так как в нем нет цели в числовом выражении. Если выручка компании повысится на 5 долл., сотрудники заявят, что задача выполнена, и прекратят прилагать усилия. И наоборот, если цель очевидно завышена и нереалистична, например «повысить выручку на 5000%», ее никто не воспримет всерьез или сотрудники вскоре сдадутся, и будь что будет. Показатели должны быть достижимыми, но при определенных усилиях.

KPI должны быть прозрачными

По крайней мере для тех, кто отвечает за их выполнение, а лучше и для всех остальных. Сотрудники должны получать обратную связь и четко понимать, приносят ли их усилия результаты или им лучше что-то изменить в своей деятельности. Стратегические показатели и ключевые показатели эффективности в компании Warby Parker доводятся до сведения всех сотрудников и регулярно

[1] URL: http://www.oreilly.com/data/free/building-data-science-teams.csp.

(хотя бы раз в квартал) обсуждаются со всем персоналом во время общих собраний рабочего коллектива.

KPI должны отражать цели, которых хочет добиться компания

Легко попасться в ловушку и начать отслеживать то, что легко измерить, например время ответа на телефонные звонки в центре обслуживания клиентов, когда истинная цель может заключаться в том, чтобы повысить степень удовлетворенности клиентов. Как гласит афоризм, «мы придаем важность тому, что способны измерить»[1]. Для этого могут потребоваться новые процессы сбора данных и оценки эффективности. Проводите дополнительную работу и меняйте то, что вы действительно стремитесь изменить.

Как и цели, KPI должны соответствовать критериям SMART[2] и быть:

— конкретными (**S**pecific);

— измеримыми (**M**easurable);

— достижимыми (**A**chievable);

— ориентированными на результат (**R**esult-oriented);

— ограниченными во времени (**T**ime-bound).

Возможно, они должны быть даже SMART*ER* за счет добавления еще двух критериев: «подвергаться оценке» (**E**valuated) и «подвергаться обзору/вознаграждаться» (**R**eviewed/**R**ewarded).

ПРИМЕРЫ КЛЮЧЕВЫХ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ

Бернард Марр[3] выделил 75 общих ключевых показателей эффективности[4]. Они включают такие области, как финансовая деятельность и понимание клиентов (табл. 6.1).

Таблица 6.1. Набор стандартных KPI для бизнеса по версии Бернарда Марра

[1] Feinberg R. A., Kim I-S., Hokama L., de Ruyter K. and Keen C. Operational deteminants of caller satisfaction in the call center. *Int. J. Service Industry Management* 11, no. 2 (2000): 131–141.

[2] URL: https://en.wikipedia.org/wiki/SMART_criteria.

[3] URL: https://www.linkedin.com/pulse/20130905053105-64875646-the-75-kpis-every-manager-needs-to-know.

[4] Marr B. Key Performance Indicators (KPI): The 75 measures every manager needs to know. London: Financial Times Press, 2012.

Финансовая деятельность	Понимание покупателей
Чистая прибыль Коэффициент доходности Коэффициент валовой прибыли Чистая прибыль от основной деятельности Прибыль до уплаты налогов, процентов, износа и амортизации (EBITDA) Рост выручки Совокупная прибыль акционеров (TSR) Добавленная экономическая стоимость (EVA) ROI Рентабельность привлеченного капитала (ROCE) Коэффициент рентабельности активов (ROA) Рентабельность собственного капитала (ROE) Соотношение собственных и заемных средств Цикл обращения денежных средств (CCC) Коэффициент оборотного капитала Коэффициент операционных расходов (OER) Соотношение капитальных затрат и объема продаж Коэффициент ценности акции (P/E ratio)	Индекс потребительской лояльности (NPS) Коэффициент удержания клиентов Индекс удовлетворенности потребителей Показатель доходности клиента Пожизненная ценность клиента (CLV) Показатель возвращаемости клиентов Вовлеченность клиентов Жалобы клиентов

Тем не менее каждая компания должна выбрать и скорректировать под себя собственный набор KPI, учитывающий область деятельности, конкретную бизнес-модель, этап жизненного цикла компании и ее особые цели и задачи. Например, стратегические показатели и KPI компании Warby Parker практически не пересекаются с перечисленными в списке Марра. Со списком все в порядке, он охватывает большинство стандартных бизнесов и их потребностей, просто он не учитывает, что каждая компания уникальна.

У компании Warby Parker серьезная социальная миссия: на каждую проданную пару очков мы отдаем пару очков тем, кто в них нуждается. Поэтому неудивительно, что наши стратегические цели и KPI связаны с благотворительной программой Do Good, потому что именно на ее дальнейшем продвижении мы хотим сконцентрироваться. Мы разрабатываем и производим собственные модели очков, так что у нас есть KPI, ориентированные на улучшение этого направления бизнеса.

Основная мысль, которую я хочу до вас донести, в том, что нет и не может быть единого готового набора KPI для всех без исключения. Для их разработки топ-менеджмент компании должен тщательно обдумать,

в каком направлении она должна развиваться, а для их выполнения всему персоналу компании следует прилагать серьезные усилия на протяжении следующего года.

Система сбалансированных показателей, предложенная Р. Капланом и Д. Нортоном[1], пытается обеспечить, чтобы набор KPI давал целостную картину деятельности компании в четырех областях: финансовой, в работе с клиентами, во внутренних бизнес-процессах, а также в обучении и развитии. Они сравнили управление компанией с управлением самолетом[2]. Чтобы поднять самолет в воздух и долететь до пункта назначения, пилоту нужно *одновременно* контролировать запас топлива, скорость полета, координаты маршрута, внешние условия и так далее. Невозможно в одном полете сосредоточиться исключительно на уровне топлива, а в следующем полете думать только о координатах маршрута. Все эти компоненты нужно рассматривать как единую стратегию.

Если вы зайдете в кабину пилота, то увидите десятки, если не сотни, датчиков, измерительных приборов и рычагов. Однако на самом деле пилот и второй пилот в штатных ситуациях, как правило, отслеживают лишь небольшой набор самых главных показателей. (Если бы вам, как мне, довелось управлять безмоторным самолетом, вы бы довольно быстро уловили, какой минимум приборов действительно необходим: альтиметр, компас, указатель скорости полета и указатель скорости набора высоты (вариометр). Все!) Компас важен. Свет на бортовой кухне важен не настолько. Вы увидите множество сигнальных ламп на панелях управления. Конечно, пилот отреагирует, если какая-то из них загорится, но в штатном режиме он может просто о них забыть. Иными словами, в компании действительно должны быть инструменты для отслеживания сотен или тысяч операционных и диагностических показателей, но сам процесс отслеживания может быть делегирован на уровень операционной деятельности. Эти панели и показатели могут быть локализованы под отдельные бизнес-подразделения или команды, но с ключевыми показателями эффективности все по-другому: этот небольшой набор показателей должен быть понятен для всех.

Итак, сколько ключевых показателей эффективности у вас должно быть?

---

[1] Kaplan R. S. and Norton D. P. The Balanced Scorecard: Translating Strategy into Action. Harvard Business Review Press, Boston: Harvard Business Preview Press, 1996.

[2] Kaplan R. S. and Norton D. P. Linking the Balanced Scorecard to Strategy, California Management Review 39, no. 1 (1996): 53–79.

СКОЛЬКО KPI ДОЛЖНО БЫТЬ?

KPI должны охватывать все основные области бизнеса и те аспекты, которым уделяется особое стратегическое внимание в этом временно́м периоде, обычно в течение года. В компании может быть четыре-пять основных направлений или заинтересованных групп, которые могут, но не должны, совпадать с топ-менеджментом компании. Например, это может быть финансовое направление, за которое отвечает коммерческий директор, или стратегические технологические цели под управлением технического директора и команды его специалистов и так далее.

Роберт Шампейн[1] полагает, что по каждому из этих направлений могут быть две-пять стратегических целей, каждая из которых может быть связана с одним-тремя KPI. При этом лучше, если общее число KPI будет в более низких значениях, рассчитанных по формуле: 5 × (2–5) × (1–3) продуктов. Он называет максимальное их количество от 20 до 30. Один из читателей ответил ему в комментариях, что «20 — это уже много». Каплан и Нортон предлагают 16–25 показателей.

Если у вас слишком много ключевых показателей эффективности, у сотрудников компании будет рассеян фокус внимания, они будут стараться выполнять несколько задач одновременно, в результате чего их эффективность может только снизиться. Например, небольшая компания не в состоянии одновременно расширить продуктовую линейку, повысить степень удовлетворенности покупателей, увеличить выручку и выйти на международный рынок. Это слишком, сотрудники выбьются из сил и будут обречены на провал. Вместо этого стоит сконцентрироваться на менее масштабном, но более целостном наборе целей, задач и KPI, которые будут понятны всем и достижимы.

ЦЕЛИ И ФОРМУЛИРОВКИ КЛЮЧЕВЫХ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ

Если ключевые показатели эффективности должны соответствовать критериям SMART, то они должны быть конкретными и измеряемыми. Это означает, что в их формулировках не должно быть общих, двусмысленных или непонятных глаголов, таких как «улучшить», «повысить», а также таких существительных и прилагательных, как «лучший», «ведущий», «качество». Стейси Барр, специалист по оценке эффективности, называет такие слова «словами-хамелеонами»[2]. Вместо

[1] URL: http://www.onvectorconsulting.com/too-many-kpis-tips-for-metrics-hoarders/.

[2] URL: http://www.staceybarr.com/measure-up/setting-your-goals-without-jargon-hbr/.

этого она рекомендует взять какую-нибудь неясную цель, например «трансформировать результативность наших клиентов», побеседовать с нужными людьми, понять смысл «слов-хамелеонов» и заменить их на более конкретную формулировку, например «когда наши клиенты работают вместе с нами, они способны быстрее достигнуть своих целевых показателей». После этого становится проще определить конкретные, измеримые показатели для достижения этой цели, например «сократить среднее время выполнения плана» или «повысить процент выполненных задач к указанной дате».

Ранее в качестве примера KPI я упоминал «удвоить число активных пользователей к концу года». Это тот случай, когда точные определения чрезвычайно важны.

Понятие «активный пользователь» можно трактовать довольно широко. В онлайновом игровом сообществе это определение может относиться к пользователям, которые просто зарегистрировались за последние 30 дней, или сыграли определенное количество игр, или потратили на игры определенное количество часов. Это определение нужно недвусмысленно уточнить в момент, когда устанавливаются показатели.

Итак, какие KPI можно отнести к хорошим, а какие — к плохим? Мария Микаллеф[1] приводит отличные примеры.

Вот хорошие цели для KPI.

— «Мы сократим количество недостающих контейнеров для бытовых отходов на 5% в следующем году».

— «Мы увеличим число наших клиентов из Италии на 20% к концу 2011 года».

В каждой из этих целей содержатся конкретные числовые показатели (при условии, что концепции «недостающих» и «клиентов» недвусмысленны или четко определены), они измеряемы и ограниченны во времени. Как насчет плохих целей?

Приведем плохие цели для KPI.

— «Мы стремимся стать лучшей транспортной компанией в регионе».

— «Мы улучшим нашу работу с жалобами клиентов».

— «Мы ответим на 75% всех жалоб в течение пяти дней».

[1] Micallef M. Key Performance Indicators for Business Excellence. URL: http://www.academia.edu/12077200/Key_Performance_Indicators_for_Business_Excellence.

Давайте проанализируем эти цели.

В первом случае вопрос очевиден: что значит «лучшей»?

Во втором случае вопрос тоже напрашивается сам собой: как «улучшим»?

А вот третья цель особенно интересна. «Ответим на 75% жалоб» — это весьма конкретно. «В течение пяти дней» — тоже ясно и с ограничением по времени. Фактически, если предположить, что эта цель достижима, то она соответствует всем пяти критериям SMART. Что же тогда не так?

Проблема в оставшихся 25% жалоб. Как быть с ними? Как говорит Мария Микаллеф, «это плохая цель, если на обработку оставшихся 25% жалоб уйдет три месяца». Одна из задач, которую вы должны держать в голове при разработке показателей, — то, что ваши сотрудники не должны осознанно или бессознательно пользоваться подобными «лазейками» в формулировках, чтобы формально выполнять поставленные перед ними задачи, но фактически не способствовать достижению стратегических целей компании[1]. В данном случае негативных отзывов от тех 25% клиентов, на чьи жалобы не отреагируют в течение пяти дней, будет достаточно, чтобы уничтожить репутацию вашей компании.

В этих двух главах мы обсудили ключевые показатели эффективности, которые определяют, чего стремится достигнуть компания и на что обращать внимание для разработки качественных диагностических и операционных показателей (какие аспекты компания собирается отслеживать и оптимизировать). Кроме того, мы поговорили о видах анализа, которые можно применять при работе с этими данными. Следующий шаг в аналитической цепочке ценности заключается в «упаковке» сделанных выводов и рекомендаций, чтобы представить их коллегам, руководству и тем людям, от которых зависит принятие решений. То есть вам необходимо рассказать историю на основе этих данных. Это тема следующей главы.

[1] См. Kerr (1975), где приводятся примеры «испорченных» мотивационных программ, и одна из причин этого — «увлечение “объективными” критериями: руководители стремились установить простые количественные стандарты, согласно которым можно было бы оценивать и вознаграждать результативность сотрудников. Подобные усилия могут быть успешными внутри компании, но, скорее всего, приведут к подмене целей, если их использовать где-то еще». URL: http://www.ou.edu/russell/UGcomp/Kerr.pdf.

ГЛАВА 7

Сторителлинг на основе данных

Когда вам удается удачно визуализировать свою мысль, собеседник моментально ее ухватывает, и диалог продолжается. Вы получаете ответную реакцию. Это повышает продуктивность. Это гораздо эффективнее, чем разговор по телефону или письмо по электронной почте. Вы сразу же доносите свою идею до многих людей.

Офер Менделевитч[1]

В предыдущих двух главах мы обсудили виды анализа, от описательного до каузального, а также вопросы разработки показателей, включая особенно важные — KPI. В этой главе мы продвинемся дальше по аналитической цепочке ценности — перейдем к обсуждению того, как «упаковывать» сделанные выводы и рекомендации и презентовать их руководству и другим заинтересованным лицам, чтобы это способствовало повышению качества дискуссии и процесса принятия решений на всех уровнях.

В этой главе приводится общий обзор процесса и целей передачи и распространения аналитических выводов в компании с управлением на основе данных: мы рассмотрим, почему и что может составлять аналитическую коммуникацию, но не будем останавливаться на том, как ее осуществлять. Я расскажу о подготовительном этапе, о чем вам стоит задуматься перед тем, как приступить к подготовке презентации или визуализации. Чтобы внести конкретику, я остановлюсь на инструменте, позволяющем подбирать графики и диаграммы, и на контрольном списке относительно визуализации данных. Надеюсь, они, а также ссылки на источники, скажут сами за себя. После этого нам останется кратко коснуться некоторых вопросов подготовки презентации, таких как общая структура и основное сообщение.

[1] Цит. по книге М. Барлоу Data Visualization: A New Language for Storytelling (O'Reilly).

Сторителлинг

«Каждый набор, каждая база данных, каждая таблица способны рассказать целую историю», — уверен Стюарт Франкел, CEO компании Narrative Science. Работа специалиста по анализу данных заключается в том, чтобы увидеть эту историю или хотя бы историю, интересную для компании, сформулировать ее и донести до аудитории. Более того, аналитикам следует позаботиться о точности истории, которая должна быть подтверждена практикой. В противном случае люди придумают свою историю, опираясь на сомнительные данные. В книге Дэвенпорта и др. Analytics at Work (с. 138–139) приводится в качестве примера случай, когда один из руководителей больницы был уверен, что главный фактор, влияющий на удовлетворенность пациентов качеством обслуживания, — качество еды. Когда аналитики взялись проверить это утверждение, оказалось, что это был один из наименее значимых факторов в наборе из еще 30. Убеждение руководителя было очень далеко от реальности. Чем объяснялось это несоответствие? Руководитель поговорил с *двумя* пациентами, которые пожаловались на качество еды. Он сделал вывод на основе случайных эпизодов, в то время как выводы аналитиков строились на основе репрезентативной выборки данных и объективного статистического анализа.

Учитывая сказанное, на бытовом уровне под историей может подразумеваться эпизод из жизни, однако что я вкладываю в этот термин в нашем контексте, то есть в рамках презентации в компании с управлением на основе данных?

Взгляните на рис. 7.1. Вам ничего не кажется необычным или интересным?

Очевидно, 2009 год для Twitter напоминал аттракцион «американские горки»: беспрецедентный рост числа подписчиков и не менее грандиозное падение (при этом все-таки наблюдалась положительная динамика и рост количества пользователей). За этой одной кривой стоит насыщенная событиями история. Первый подъем (примерно в марте 2007 года) объяснялся шумихой вокруг Twitter на ежегодной конференции South by Southwest Interactive Conference, когда на сервис впервые обратили внимание и количество его пользователей сразу утроилось. Замедление роста после второго подъема (примерно в марте 2008-го) объясняется тем, что тогда Twitter начал активно вносить в черный список спамеров. В 2009 году сервис получил уже широкую известность, в апреле на пике популярности, как раз перед падением, Эштон Кутчер поспорил с телеканалом CNN, у кого из них первым будет

один миллион подписчиков (Эштон выиграл буквально через полчаса), а Опра Уинфри первый раз отправила сообщение в Twitter и сделала это в прямом эфире. Аналогичная кривая, построенная на данных пользователей из Австралии, в чем-то похожая на кривую по США, но имеет свои отличия. Так, например, последний рост количества пользователей Twitter в Австралии в 2013 году совпал с проведением выборов на федеральном уровне.

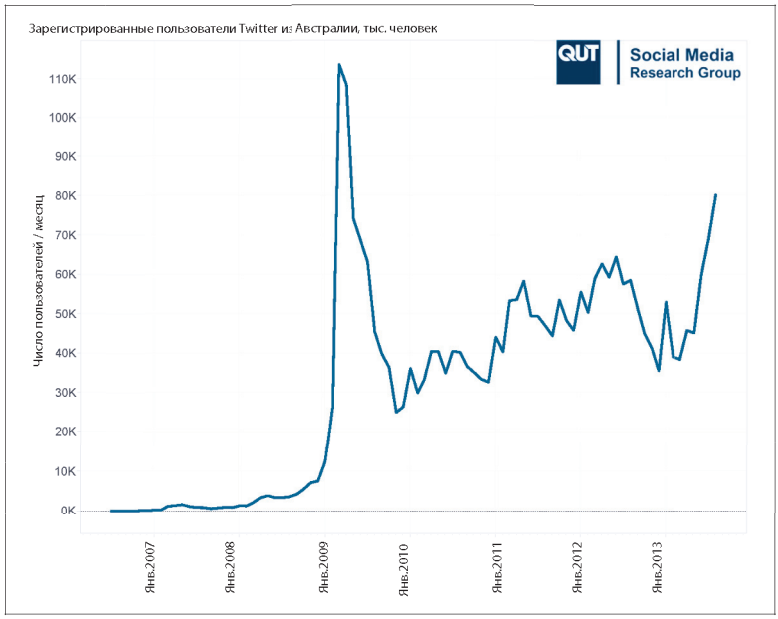


Рис. 7.1. Регистрация новых пользователей из Австралии в Twitter на протяжении времени

Источник: http://socialmedia.qut.edu.au/2014/08/04/first-steps-in-exploring-the-australian-twittersphere/

Таким образом, история должна содержать основные выводы, особенности данных или присущие им закономерности, чтобы по возможности раскрывать причины происходящего, а также смотреть в будущее, делать прогнозы и формулировать рекомендации для компании. По Стивену Фью, «визуализация данных — это применение средств визуального представления для изучения, анализа и презентации количественных данных». В данной книге я рассматриваю сторителлинг как дополнительный интерпретативный слой, повествовательную структуру на вершине визуализации данных. Рис. 7.1, дополненный

описательной частью, более полезен, чем просто рис. 7.1. График и описание дополняют друг друга. Требуется качественная визуализация, чтобы обнаружить закономерности в данных в ходе проведения анализа, а затем продемонстрировать их аудитории. И помимо этого требуется знание точной и достоверной истории для интерпретации данных и построения возможных прогнозов.

В идеале в данном случае можно включить информацию о переломных моментах в график и таким образом усилить историю и сделать более самодостаточной (рис. 7.2).

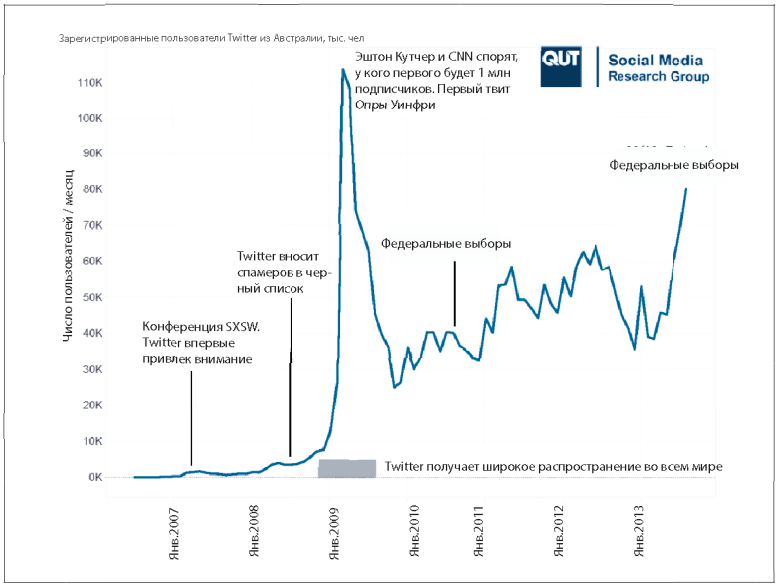


Рис. 7.2. Аннотированная версия рис. 7.1

Поиск истории и ее интерпретация включают использование ряда аналитических техник, в которые обычно входит разведочный анализ, то есть, условно говоря, визуализация данных с помощью таблиц и диаграмм (глава 5). Эта глава посвящена визуализации данных, но это не введение в тему: я бы не смог сделать это на должном уровне, кроме того, есть немало отличных книг специально по теме. Начинать я рекомендую с золотого стандарта: книг Эдварда Тафти Envisioning Information («Представление информации»), Visual Explanations («Визуальные объяснения») и The Visual Display of Quantitative Information (Graphics Press) («Визуальное отображение количественной информации»). Третья книга особенно хорошо поможет вам понять, как мыслит

дизайнер и критик. В этой книге Тафти представил важные концепции «графического мусора» и соотношение данных и чернил (Data-to-ink ratio), то есть элементов, несущих информационную нагрузку. Обе эти концепции я объясню далее.

Если вы хотите почитать что-то более практически направленное, рекомендую книги Стивена Фью Now You See It (Analytics Press), которая в большей степени сосредоточена на визуализации данных для изучения и анализа количественных данных, а также Show Me The Numbers (Analytics Press), посвященную процессу презентации. Для ознакомления с вопросами визуализации данных в виртуальном пространстве начните с книги Скотта Мюррея Interactive Data Visualization (O'Reilly). Кроме того, эту главу не стоит рассматривать как руководство по стилю. Для этих целей настоятельно рекомендую книгу Доны Вонг The Wall Street Journal Guide to Information Graphics (W. W. Norton & Company).

Первые шаги

Прежде чем размышлять над тем, как лучше всего представить данные, информацию, результаты анализа, следует ответить на три вопроса:

— Чего вы хотите добиться?

— Кто ваша аудитория?

— Каким средством вы воспользуетесь?

Это основные факторы, которые сузят выбор типа презентации, стиля, уровня технических средств, способных донести ваше сообщение максимально эффективно. Только после того, как вы определитесь с этими тремя вопросами, можно будет перейти к более практическим аспектам — как вы собираетесь структурировать содержание и оформлять его визуально.

ЧЕГО ВЫ ХОТИТЕ ДОБИТЬСЯ?

Какова ваша цель? Зачем вы делаете эту презентацию или отчет? Какого результата вы надеетесь достигнуть? Предположительно, эту цель следует определить еще до начала самого анализа, но у вас должно сложиться четкое понимание, зачем вы представляете эти данные или результаты анализа, к каким выводам вы пришли и что, по вашему мнению, произойдет дальше.

Например, если вы проводите только описательный анализ, его цель может состоять в том, чтобы читатели получили более ясное понимание системы, уловили взаимосвязи, величину и возможность изменений основных компонентов, то есть цель — поделиться знаниями. Если вы проводите анализ результатов A/B-тестирования, то его цель может заключаться в том, чтобы оценить, насколько эффективны разные варианты решения задачи по сравнению с контрольными показателями, а также уверенность в результатах и потенциальное увеличение выручки, подтверждающее реальность решения. В этом случае цель может быть в том, чтобы принять решение и обеспечить, чтобы новая характеристика или функция стала доступна всем пользователям. Эти два вида анализа отличаются методами проведения, преследуют разные цели и требуют разных стилей презентации.

Рассмотрим подробнее пример с результатами анализа A/B-тестирования. В этом случае специалист по анализу данных должен провести собственно анализ, прийти к выводу относительно значения и достоверности результатов и предложить свои рекомендации: надо ли внедрять эту характеристику в массовое производство. В своей презентации он должен отразить рекомендации и привести подтверждения: так мы проводили тестирование, это показатели, вызывающие интерес, вот что мы обнаружили, это небольшая неясность, с которой мы столкнулись, а вот почему мы пришли к финальному заключению.

КТО ВАША АУДИТОРИЯ?

Следующий вопрос, на который нужно ответить, касается аудитории, для которой готовится презентация. Насколько хорошо эти люди подкованы технически, умеют ли они оперировать данными? Каковы их ожидания? Каковы их уровни заинтересованности и мотивации? Насколько они заняты? В некотором смысле аналитик должен уметь добиваться своих целей *вопреки* аудитории. Тема презентации — это, возможно, главная задача, на которой он сосредоточен в последние дни или недели. Но для слушателей презентации это может быть лишь одним из десяти решений, которые они приняли сегодня, особенно когда речь идет о топ-менеджменте компании. У аналитика должно быть четкое понимание статистических техник, которые он применял в работе, в то время как аудитория, скорее всего, не имеет об этом представления. Аналитик поглощен цифрами, кодами, статистикой, тогда как слушателей волнует только необходимость принятия бизнес-решений и последующий эффект. При подготовке презентации

аналитик должен принять во внимание все перечисленные факторы и структурировать материал так, чтобы добиться максимальной результативности.

Например, если вы понимаете, что на разговор с большим боссом вам отведут всего несколько минут, будьте лаконичны и конкретны: «Я рекомендую предпринять следующие меры, так как они позволят нам получить миллион дополнительного дохода в течение следующего года». В других случаях, например в часовой презентации для других специалистов по статистике, можно максимально углубиться в технические детали. Возможно, их заинтересуют степени свободы, доверительные интервалы, графики плотности распределения и другие аспекты.

Финансовые директора обычно чувствуют себя комфортно при работе с большими таблицами финансовых показателей (можно ли утверждать, что эта форма получения информации для них предпочтительна — уже другой вопрос). Для более широкой аудитории, например во время общего собрания, лучше облегчить информацию и представить общие выводы без технических подробностей. Решите, какой способ представления данных подходит вам больше всего, и структурируйте материал соответственно.

КАКИМ СРЕДСТВОМ ВЫ ВОСПОЛЬЗУЕТЕСЬ?

Наконец, определитесь со средством: будет ли это доклад в письменной форме, графическая презентация, например в PowerPoint, дашборд или инфографика.

Частично этот вопрос связан с предыдущим. Например, если вы выступаете на общем собрании, у вас есть выбор между графической презентацией или устным докладом. Для финансового директора лучше подготовить письменный отчет и включить в него необходимые таблицы и графики по тем направлениям, которые ему нужны и которые он ожидает увидеть. Для выступления перед руководителями нескольких направлений, возможно, вам понадобится подготовить презентацию в PowerPoint.

Решение относительно средства презентации в совокупности с пониманием общего уровня заинтересованности аудитории и объема времени, которое будет отводиться на презентацию, поможет определить, насколько глубокой она должна быть. Если у вас только три минуты, чтобы выступить перед топ-менеджером, то презентация в PowerPoint на 37 слайдов с кучей технических деталей точно не понадобится.

Конечно, можно остановить свой выбор на презентации в PowerPoint, но тогда это будут два-три слайда. Еще один важный момент: не стоит копировать визуальную информацию из одного средства и использовать ее для другого. Например, копирование большой таблицы из письменного отчета и размещение ее на слайде в PowerPoint, который вы собрались демонстрировать на общем собрании, будет малоэффективным. Нужно подогнать каждый слайд, график или таблицу под то средство, которым вы хотите воспользоваться.

ПРОДАВАЙТЕ!

Качественно спланированный эксперимент, тщательно отобранные показатели и, самое важное, четко заданный вопрос обеспечивают наибольшую вероятность обнаружить доминирующие закономерности в данных и найти ответы на поставленные вопросы. Работа аналитика состоит в том, чтобы найти и проиллюстрировать самые очевидные и наиболее подходящие закономерности, интерпретировать их и транслировать с точки зрения влияния на бизнес. Однако это все-таки будет лишь одной интерпретацией данных из возможных. На основе этих же данных другие сотрудники могут прийти к другим заключениям. Именно поэтому эксперт в области визуализации данных Себастьян Гутьеррес сравнивает аналитика, презентующего данные с помощью визуализации, с продавцом: «Вы пытаетесь продать какую-то идею: мы должны увеличить бюджет, мы должны изменить базу данных, мы должны привлечь больше пользователей… У вас есть сообщение, которое вы стремитесь донести. Когда я представляю данные неспециалистам в этой области, то отношусь к этому как к упражнению по маркетингу».

Что вы продаете? По крайней мере, две вещи. Во-первых, если есть несколько интерпретаций, задача аналитика— выбрать и продвинуть наиболее объективную, логичную и *экономичную* (простую) из них, а также суметь обосновать свою позицию. Во-вторых, если аналитик затратил столько усилий на сбор данных, их обработку, анализ, возможно, построение модели и в итоге обнаружил нечто действительно важное, что способно оказать влияние на развитие бизнеса, он изо всех сил будет стремиться к тому, чтобы результаты его работы были применены на практике. Аналитик старается продать *действие* (что следует сделать) и *результат* (что получится в итоге этого действия). Мы еще вернемся к этому моменту в главе 9. Иными словами, специалист по анализу данных не пассивный транслятор данных, информации, выводов — он должен активно продавать эти идеи.

Более того, Себастьян отмечает, что, когда аналитик подходит к этому процессу с позиции маркетинга и у него есть идея, которую он должен продвинуть, это стимулирует его искать больше данных, чтобы получить более убедительную и подтвержденную фактами историю. Важно, что корпоративная культура организации должна стимулировать аналитика, чтобы он стремился оказать максимальное влияние на деятельность компании. Кен Рудин, руководитель аналитического направления в Facebook, а до этого в компании Zynga, подтверждает это примером:

> Смысл аналитики в оказании влияния… В нашей компании [Zynga], если вы провели блестящее исследование и сделали потрясающие выводы, но ничего не изменилось, результативность вашей работы равна нулю.

Визуализация данных

Теперь, когда мы имеем более ясное представление о том, что такое сторителлинг, а также о роли аналитика и его мотивации, давайте обсудим некоторые технические аспекты визуализации данных. Как уже упоминалось в начале этой главы, наше обсуждение не будет полноценным руководством по этой теме. Я остановлюсь на нескольких ключевых моментах и свяжу их с общими комментариями, типичными ошибками и да, с тем, что больше всего раздражает лично меня.

Итак, предположим, что аналитик выбрал правильные метрики, правильные измерения (например, систематизировал данные по месяцам или по каналам продаж), обнаружил интересные и значимые закономерности в этих данных, Следующий шаг, который он должен предпринять, — выбрать форму презентации этих данных. В некоторых случаях это может быть таблица, но чаще всего останавливаются на диаграмме.

ВЫБОР ДИАГРАММЫ

У аналитика большой выбор разных типов диаграмм. Подходящий тип диаграммы или визуализации зависит от типа переменных (непрерывные, дискретные, категориальные или порядковые), от того, сколько переменных или факторов требуется включить в диаграмму, и даже от значений переменных. Например, составная столбиковая диаграмма способна справиться с двумя категориями данных, но не с бóльшим числом (рис. 7.3).

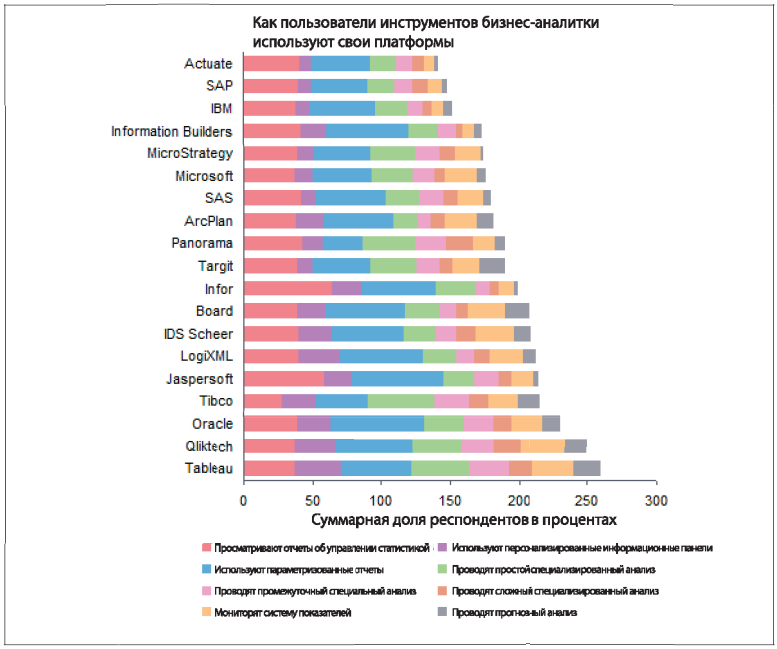


Рис. 7.3. Пример составной столбиковой диаграммы (показывающей, как пользователи инструментов бизнес-аналитики используют эти продукты) с относительно большим числом категорий (восемь). Легче всего между платформами сравнить крайнюю левую категорию, так как она выровнена по оси y. Однако интерпретировать результаты по другим категориям не так просто, поскольку они отличаются по ширине и расположению. Например, как сравнить между платформами крайнюю правую категорию?

Источник: Джон Пелтир (http://peltiertech.com/stacked-bar-chart-alternatives/)

Для сравнения: рис. 7.4 содержит те же самые данные, но их легче сравнить между платформами, хотя и за счет потери понимания суммарной доли респондентов в процентах (то есть полной ширины столбца на рис. 7.3).

Выбор типа диаграммы — основной фактор с точки зрения способности сделать презентацию данных понятной для пользователей. Так на чем же остановить свой выбор в условиях такого разнообразия? Один из способов — сосредоточиться на одной из четырех причин, по которым мы вообще строим диаграмму.

Сравнение

Например, сравнение групп или сравнение изменений во времени.

Распределение

Необходимость показать изменчивость набора данных.

Взаимосвязи

Необходимость отразить корреляцию или взаимосвязь между переменными.

Сравнение

Необходимость показать, как распределяются данные между двумя или более категориями.

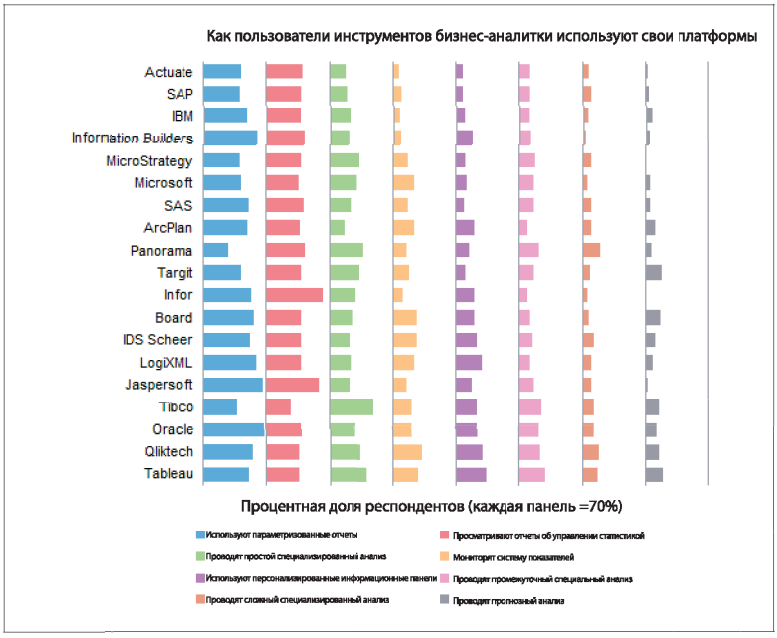


Рис. 7.4. Те же самые данные, что и на рис. 7.3, представлены в виде панельной диаграммы. В этом случае гораздо проще интерпретировать сравнение между категориями.

Источник: Джон Пелтир (http://peltiertech.com/stacked-bar-chart-alternatives)

На рис. 7.5 приведены примеры разных типов диаграмм и то, как они соотносятся с выделенными нами четырьмя целями. Мы выбрали наиболее распространенные типы диаграмм, хотя существует еще множество других. Например, здесь никак не охвачены данные из социальных сетей или геопространственные данные.

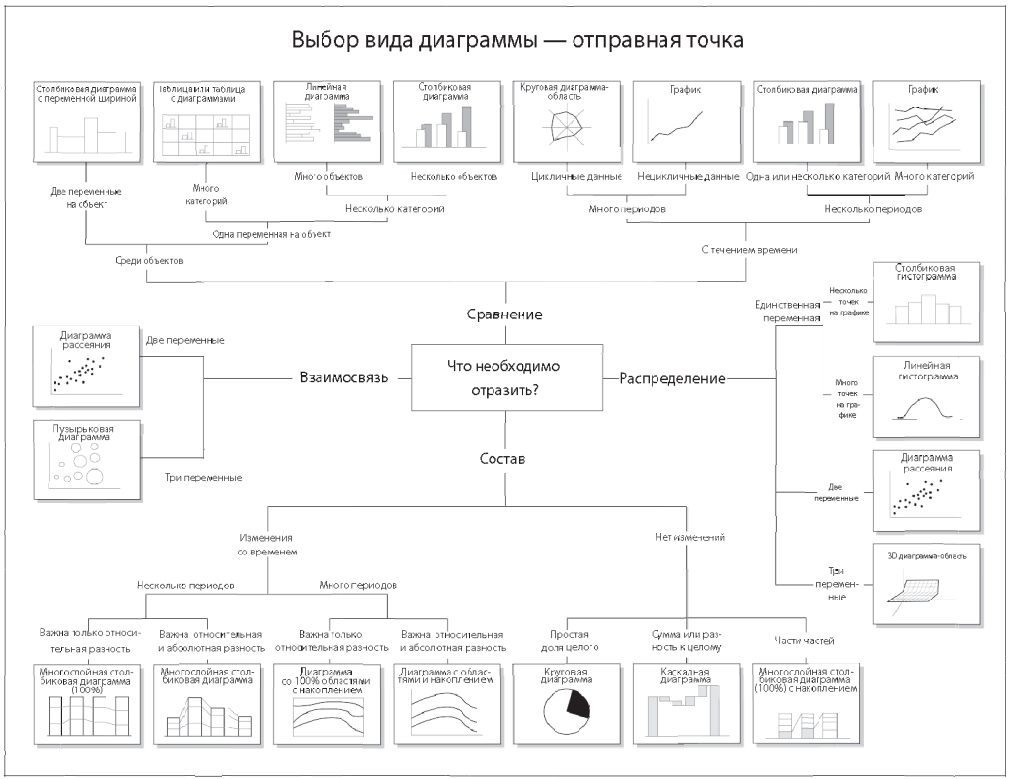


Рис. 7.5. Существует много разных типов диаграмм, каждый из которых отвечает определенной задаче. Выберите тот тип, который оптимально подходит для решения вашей задачи

Источник: Эндрю Абела (http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html). Воспроизводится с разрешения

Более полное представление типов диаграмм можно найти в виде инфографического постера Graphic Continuum[1], но, к сожалению, он слишком масштабный и детальный, и его невозможно без потери качества разместить на одной книжной странице. Кроме того, я рекомендовал бы изучить галерею визуализации D3[2]. D3 — это популярная библиотека JavaScript, которой можно воспользоваться для выполнения более интересной, интерактивной или специализированной визуализации данных.

Как вы сами видите, для работы с конкретным набором данных можно использовать разные типы диаграмм, в каждой из которых будет делаться акцент на разных характеристиках данных. Главное — пробовать разные варианты. Исследуйте «дизайнерское пространство» в поисках средств, которые помогут лучше всего рассказать вашу историю, но при этом не лишат ее достоверности и объективности (например, не усекайте ось *y*, чтобы исказить угол наклона в линейном графике[3]).

ВЫБОР ЭЛЕМЕНТОВ ДИАГРАММЫ

Выбор типа диаграммы — относительно простая задача, так как он ограничен (хотя даже это не мешает некоторым выбирать неподходящие варианты). Но это только начало. Далее приводится контрольный список тех элементов, на которые стоит обратить внимание при построении диаграммы. Мы не будем подробно разбирать каждый из указанных пунктов, так как это не входит в задачи этой книги. Скорее, это подсказка для вас, с чего можно начать. Если вы хотите получить более глубокие знания, я вновь рекомендую обратиться к тем книгам, которые я перечислял в начале главы. Многие из элементов этого контрольного списка могут показаться очевидными; тем удивительнее, сколько встречается диаграмм, построенных с нарушением одного или нескольких из этих критериев, что не может не сказаться на их эффективности.

КОНТРОЛЬНЫЙ СПИСОК ДЛЯ ВИЗУАЛИЗАЦИИ ДАННЫХ

Визуализация данных включает множество элементов, каждый из которых требует пристального внимания. Один неверный выбор, например цвета с малым контрастом, мелкий шрифт, неподходящий тип диаграммы — и все визуальное представление испорчено. Далее приводятся

[1] URL: http://www.scribblelive.com/blog/2014/10/01/graphic-continuum.

[2] URL: https://github.com/d3/d3/wiki/Gallery.

[3] Fox J. The Rise of the Y-Axis-Zero Fundamentalists, December 14, 2014. URL: https://byjustinfox.com/2014/12/14/the-rise-of-the-y-axis-zero-fundamentalists/.

элементы полезного контрольного списка Стефани Эвергрин. В полной версии списка можно найти подробное описание каждого пункта.

Текст	Описательный заголовок из 6–12 слов в левом верхнем углу с выравниванием по левому краю. Подзаголовок и/или примечания с дополнительной информацией. Размер текста многоуровневый и читаемый. Расположение текста горизонтальное. Данные с ярлыками. Ярлыки применяются умеренно.
Выравнивание	Пропорции соблюдены. Данные выровнены. Расстояния между осями равноудаленные. График двухмерный. Минимум украшательств.
Цвет	Выбор цвета преднамеренный. Цвет применяется для выделения основных закономерностей. Цвет понятен при распечатке в черно-белом варианте. Цвет понятен для людей с проблемами с цветовосприятием. Текст достаточно контрастирует с фоном.
Линии	Линии сетки (если есть) скрыты. У графика нет рамки. На осях нет ненужных отметок. На графике одна горизонтальная и одна вертикальная ось.
Общие комментарии	График подчеркивает значимые результаты или выводы. Тип графика соответствует данным. Присутствуют данные для сравнения или обеспечения контекста. Отдельные элементы диаграммы работают вместе для усиления основного сообщения.

Фокусировка сообщения

Цель создания презентации — четко донести свое сообщение до аудитории. Для этого в вашем арсенале имеется целый ряд средств: шрифты, линии сетки, ориентация страницы. Еще одно средство — выделение цветом. Один из способов сделать сообщение сфокусированным — показывать только данные, представляющие интерес. К сожалению, иногда это может привести к отрыву от контекста. Например, предположим, что, согласно графику, Япония производила 260 тераватт-час энергии в 2009 году. Этого много или мало? Я понятия не имею. Зато все сразу становится ясно, если оставить эти данные в контексте,

но выделить цветом (рис. 7.6). Мы сразу же увидим показатели, касающиеся Японии, благодаря выделению названия жирным шрифтом и более светлому цвету столбца диаграммы. А благодаря дополнительным данным относительно других стран можно интерпретировать данные о Японии: ее уровень производства электроэнергии был высоким, но составил $^1/_3$ от уровня производства США.

Рис. 7.6. Пример эффективного использования выделения цветом. При представлении данных о Японии название страны выделено жирным шрифтом, а столбец диаграммы обозначен более светлым цветом. Это позволяет сфокусироваться на данных относительно Японии, которые, тем не менее, остаются в контексте

Источник: http://theeconomist.tumblr.com/post/3880075172/daily-chart-the-worlds-largest-nuclear-energy

Это удачный пример, как при помощи цветового выделения можно усилить сообщение. Рассмотрим противоположный случай. Следует избегать того, что Стефани Эвергрин назвала «синдром Марты Стюарт»[1], то есть чрезмерного украшательства диаграммы. Все должно быть просто. Исключите «графический мусор» и излишества и сконцентрируйтесь на данных и сообщении.

Термин «графический мусор» ввел в употребление Эдвард Тафти для обозначения элементов, отвлекающих внимание. «Графический мусор» — все визуальные элементы диаграмм и графиков, в которых нет

[1] Марта Стюарт (р. 1941) — американская телеведущая и писательница, получившая известность и ставшая успешной благодаря советам по домоводству. *Прим. перев.*

необходимости для понимания представленной информации или которые отвлекают от нее. Минималистский подход Тафти отличается категоричностью. Я предпочитаю более умеренное и прагматичное определение Роберта Косары — «любой элемент диаграммы, который не способствует прояснению сообщения»[1]. Косара признает, что в некоторых случаях может быть необходимо внести дополнительные элементы в диаграмму для выделения специфических компонентов, чтобы усилить основное сообщение или историю.

На этом этапе во многих книгах по визуализации данных (в том числе и Эдварда Тафти) для иллюстрации «графического мусора»[2] приведены диаграммы и графики, взятые из USA Today. Я не буду этого делать, а остановлюсь на новом золотом стандарте — слайды программы PRISM Агентства национальной безопасности США (рис. 7.7).

На рис. 7.7 представлена хронологическая шкала, когда разные технологические компании присоединились к программе АНБ по массовому негласному сбору информации. Это основное сообщение, но из-за множества дополнительных графических элементов внимание от него отвлечено. В верхней части слайда беспорядочно размещены 11 логотипов. Они соотносятся с желтыми овалами, но не в пропорции 1 : 1 (желтых овалов всего девять). Они только отвлекают внимание пользователя. Кроме того, на слайде размещены логотип самой программы и подразделения АНБ. Более того, на нем есть еще и зеленая стрелка. Какова ее роль? Почему данные расположены по возрастающей? Это все «графический мусор».

Подобные украшательства отвлекают внимание от основного сообщения по двум причинам:

— пользователь тратит время на рассматривание и обдумывание других элементов;

— пользователю сложно определиться, на чем сосредоточить внимание.

На рис. 7.8 приведен один из возможных вариантов исправления этого слайда. Автор слайда — Эмилэнд де Куббер. На слайде условно выделены два важных блока данных: компании и время их присоединения к программе. Девять компаний — девять логотипов.

---

[1] URL: https://eagereyes.org/blog/2013/definition-chart-junk.

[2] Поищите картинки в Google по ключевой фразе «графический мусор», и вы увидите множество примеров из USA Today. К сожалению, аналитическая колонка New York Times Magazine тоже полна вопиющими примерами.

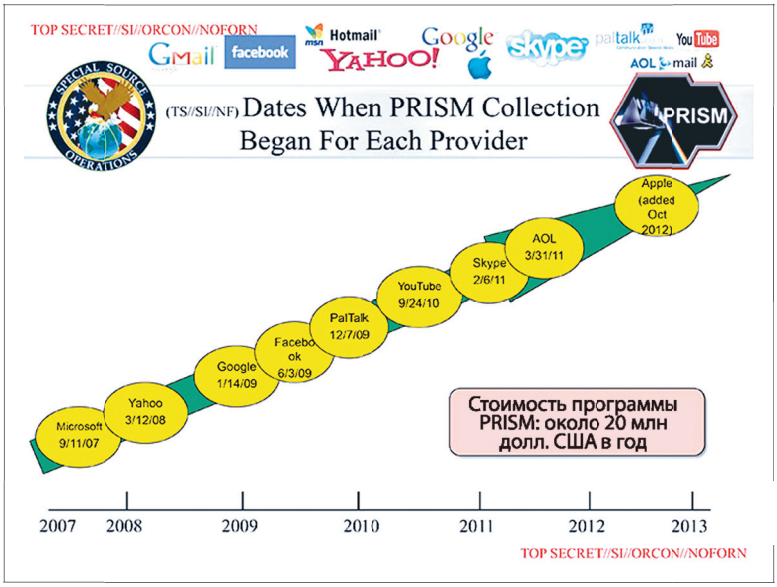


Рис. 7.7. Слайд программы PRISM АНБ США, переполненный «графическим мусором»

Источник: https://www.theguardian.com/world/interactive/2013/nov/01/prism-slides-nsa-document

Рис. 7.8. Вариант слайда, предложенный Эмилэндом де Куббером

Источник: https://www.slideshare.net/EmilandDC/dear-nsa-let-me-take-care-ou

Можно почти моментально уловить общую картину и посчитать количество компаний за каждый из указанных периодов времени (1, 1, 3, 1, 2, 1). А бросив второй взгляд на слайд, можно сосредоточиться на логотипах и понять, о каких именно компаниях идет речь. Этот вариант не идеален, но визуально информация представлена на нем более эффективно, чем на оригинальном слайде.

Организация данных

То, как будет организовано представление информации на диаграмме, зависит от выбора диаграммы, и наоборот. В рамках ограничений, которые накладывает выбор диаграммы, по-прежнему остается важным структурный выбор, например, как расположить столбцы диаграммы — горизонтально или вертикально. Самое удивительное, что даже на этом уровне есть небольшие вариации в том, как можно представить данные, так что это существенно повлияет на сообщение.

На рис. 7.9 показан среднегодовой размер оплаты труда госслужащих в Великобритании по тарифным разрядам и с делением по гендерному признаку.

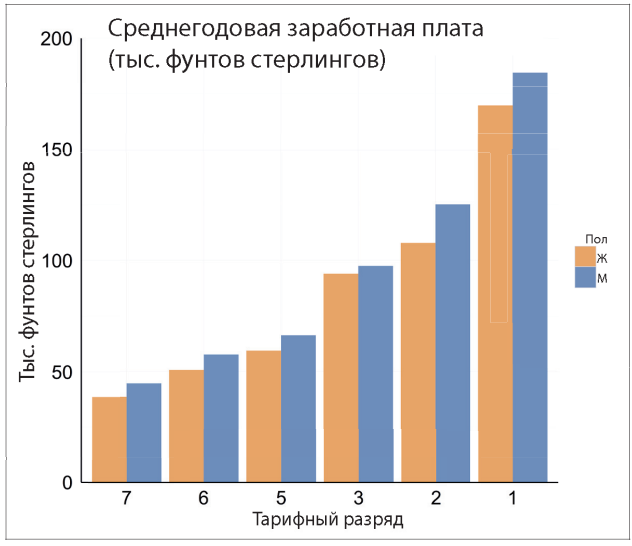


Рис. 7.9. Среднегодовая заработная плата (в тыс. фунтов стерлингов) госслужащих в Великобритании по тарифным разрядам (более низкая цифра разряда означает более высокую должность) и с делением по гендерному признаку

Источник: http://news.bbc.co.uk/2/hi/business/8044720.stm

С диаграммой все в порядке. У нее понятное название и обозначения осей. По оси *x* представлены тарифные разряды по возрастающей слева направо, как и следовало ожидать, учитывая, что в западной традиции принято направление чтения слева направо (хотя несколько вводит в заблуждение, что номера тарифных разрядов, наоборот, уменьшаются в порядке значимости). Ось *y* тоже нареканий не вызывает. Нет усечения по вертикальной оси. Интервал в 25 тыс. фунтов стерлингов кажется оправданным. При составлении диаграммы был богатый выбор цветовой палитры.

В итоге выбрали основной голубой цвет (который обычно ассоциируется с мужским полом) и дополнительный оранжевый для обозначения женского пола. Выбор вполне обоснован. В этой диаграмме нет грубых ошибок.

А теперь посмотрите, что получится, если во всех тарифных разрядах поменять местами столбцы, обозначающие пол (рис. 7.10).

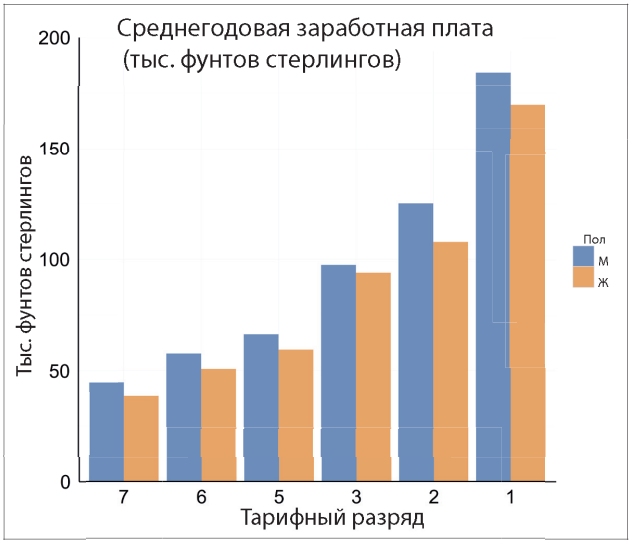


Рис. 7.10. Та же самая диаграмма, что и на рис. 7.9, за исключением того, что во всех тарифных разрядах поменяли местами столбцы, обозначающие пол. Вам не кажется, что неравенство в заработной плате по гендерному признаку бросается в глаза сильнее?

Удивительная разница. Те же самые данные, те же самые оси, те же самые интервалы и цветовая схема. Всего одно небольшое изменение кардинальным образом меняет восприятие неравенства в оплате труда

у мужчин и женщин[1]. Основное сообщение, о неравенстве оплаты труда, становится гораздо более наглядным. Первая диаграмма построена правильно, просто вторая — более наглядная.

Думаю, из этого примера очевидно, что каждая диаграмма, которую вы строите, требует индивидуального подхода. К тому же необходимо развивать в себе критическое восприятие. Этот навык приходит с практикой, в процессе работы со случаями, подобными этому. Поэтому всем специалистам по работе с данными я настоятельно рекомендую ознакомиться с книгами, которые я упоминал в начале этой главы, изучить метод trifecta checkup Кайзера Фанга — метод проверки диаграмм на наличие «графического мусора»[2], а также посещать семинары по визуализации данных и, самое главное, практиковаться. Изучайте диаграммы из Wall Street Journal, New York Times и The Economist — все они задают очень высокую планку качества. Что делает их такими эффективными и где у них бывают проколы? (Да, такое тоже случается.) Сравните диаграммы в /r/dataisbeautiful/[3] и r/dataisugly[4]. Почему первые такие ясные, а вторые такие бестолковые? Спросите себя, что бы вы сделали иначе.

Подача данных

В этом разделе мы поговорим о способах подачи сделанных выводов. Во-первых, кратко остановимся на инфографике, которая в последнее время пользуется особенной популярностью у специалистов по маркетингу. Во-вторых, изучим гораздо более важную тему дашбордов. Как уже говорилось в начале книги, многие компании считают, что у них развито управление на основе данных, просто потому что их сотрудники пользуются множеством дашбордов. Дашборды и отчеты о состоянии работ, несомненно, стали полезным и одним из наиболее распространенных инструментов. Мы рассмотрим несколько типов дашбордов и обсудим их пользу (или отсутствие таковой) для процесса принятия решений.

---

[1] Как объясняет Стивен Фью, человеческий мозг во всем стремится находить закономерности. Кроме того, мы предпочитаем более простые и плавные кривые. С точки зрения вычислений они легче поддаются расшифровке. Второй вариант, отличающийся ступенчатостью, требует больше внимания, так как мозг затрачивает больше усилий на обработку информации о форме диаграммы.

[2] URL: http://junkcharts.typepad.com/junk_charts/junk-charts-trifecta-checkup-the-definitive-guide.html.

[3] URL: https://www.reddit.com/r/dataisbeautiful/.

[4] URL: https://www.reddit.com/r/dataisugly/.

ИНФОГРАФИКА

В контексте управления на основе данных я не большой поклонник инфографики: сегодня инфографика превратилась в «веселые картинки», приправленные парой фактов, которые обычно создают дизайнеры, а не аналитики. По моему мнению, у подобной инфографики слишком низкое соотношение данных и чернил (data-to-ink ratio), как его определил Эдвард Тафти. Фактически в большинстве случаев инфографика страдает от «графического мусора» и от недостатка данных. Например, на рис. 7.11 в забавной и визуально привлекательной форме представлен размер мозга у животных с разной массой тела.

При этом более лаконичной и эффективной формой для представления этих данных могла бы стать столбиковая диаграмма или таблица:

Животное	Масса мозга в граммах (фунтах)
Кашалот	7800 (17,2)
Дельфин	1600 (3,5)
Человек (взрослый)	1400 (3)
…	…
Лягушка	0,24 (0,008 унции)

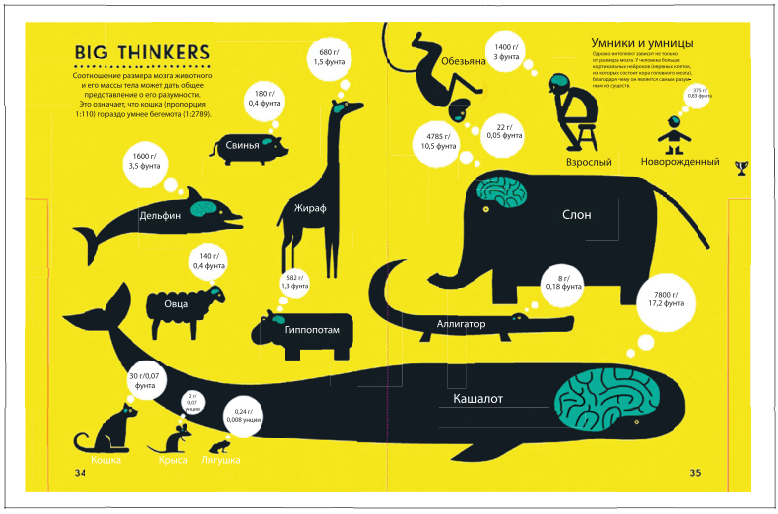


Рис. 7.11. Инфографика Big Thinkers из книги Роджерса и Блечмана (2014) Information Graphics: Animal Kingdom. Big Picture Press

На самом деле интересно здесь другое — отношение массы мозга к общей массе тела. Диаграмма, отражающая это соотношение, содержит

одно из удивительнейших открытий сравнительной биологии — закон масштаба. На рис. 7.12 показано, что масса мозга относительно общей массы тела уменьшается с увеличением массы тела[1].

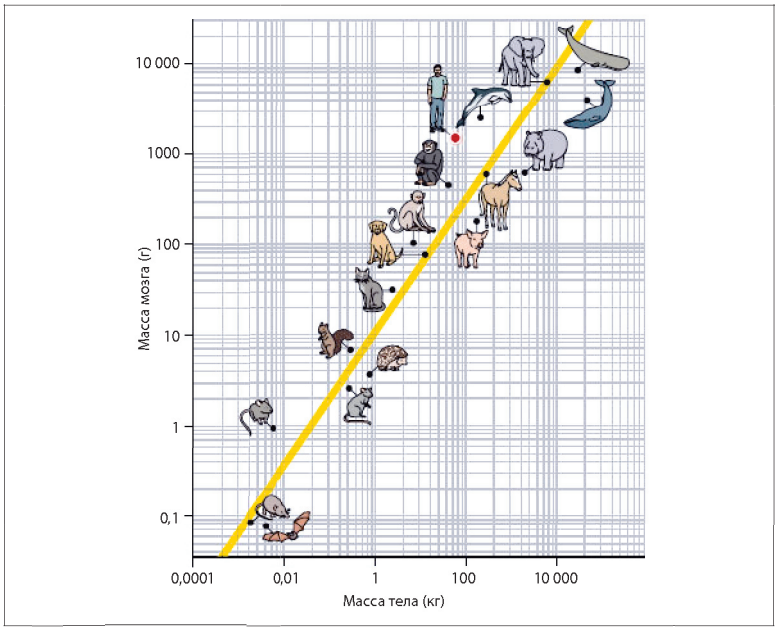


Рис. 7.12. Соотношение массы мозга и общей массы тела. (Обратите внимание: обе оси логарифмические, но интервал по оси x составляет 100x, а интервал по оси y — только 10x).

Источник: Dongen P. A. M. 1998. Brain Size in Vertebrates. Из книги The Central Nervous System of Vertebrates, Vol 3. Ed. by R. Nieuwenhuys et al., Springer

Я намеренно выбрал такой пример для иллюстрации своей мысли. Это инфографика из книги для детей, поэтому ее задача — быть увлекательной, информативной и запоминающейся. Она отлично с этим справилась. Однако когда речь заходит о компании с управлением на основе данных, такая инфографика будет бесполезна для внутреннего использования и для процесса принятия решений. Я не отрицаю,

[1] Обе оси логарифмические. Это не очевидно на первый взгляд, но интервал по оси *x* составляет 100*x*, в то время как интервал по оси *y* — только 10*x*, так что кривая графика очень крутая. Возьмем белку. У нее соотношение: 10 г масса мозга / 1 кг масса тела. Обратите внимание на человека и дельфина — оба отстоят от кривой графика: они отличаются относительно большой массой мозга *для их общей массы тела*, но все равно меньше (~5*x*), чем у мыши.

что в некоторых случаях выбор инфографики может оказаться оправданным. Недавно моя команда представила в виде инфографики наши результаты за год. Мы показывали ее на общем собрании сотрудников. Аудитория была разнообразной и преимущественно не технической, а наша цель состояла в том, чтобы быстро пройтись по наиболее важным событиям года. Так что в этой ситуации формат инфографики был уместен. Также уместен он может быть для внешней коммуникации с широкой публикой.

Интересно, что, согласно результатам последних исследований, «графический мусор», пиктограммы, цвет и контраст делают диаграммы запоминающимися[1]. И всеми этими элементами изобилует инфографика. Тем не менее еще раз повторю свою основную мысль: цель визуализации данных — стимулировать коммуникацию, ведущую к конкретным действиям. Руководителям требуется информация высокого качества, чтобы они могли не только запомнить основную мысль, но и оценить ее и убедиться, что решение, которое они собираются принять, правильное.

Пользователь должен быстро и без усилий увидеть те центральные пункты, которые отражают представленные данные, а «графический мусор» этому препятствует.

ДАШБОРДЫ

Многие компании ошибочно измеряют степень управления на основе данных количеством производимых ими отчетов и числом дашбордов, которыми они пользуются. Дашборды очень полезны и могут поддержать ряд видов деятельности, например обеспечить интерфейс для сбора данных, составления специализированных отчетов, оповещений, а также отобразить в удобном виде прогнозы и прогнозные модели. Дашборды можно условно разбить на три категории:

— управленческие или стратегические;

— аналитические;

— операционные.

Стратегические дашборды (рис. 7.13) обеспечивают общий обзор деятельности компании и, как правило, концентрируются на системе показателей (например, KPI и их цели). Дашборд должен просто и быстро

[1] URL: http://cvcl.mit.edu/papers/Borkin_etal_MemorableVisualization_TVCG2013.pdf.

помочь увидеть, достигает ли компания поставленных целей и есть ли у руководства поводы для беспокойства. Иными словами, она должна держать руку на пульсе компании и показывать обзорную картинку с высоты 15 км. В основном стратегическими дашбордами пользуется высшее руководство компании, но в компании с управлением на основе данных доступ к этим инструментам есть у более широкой аудитории.

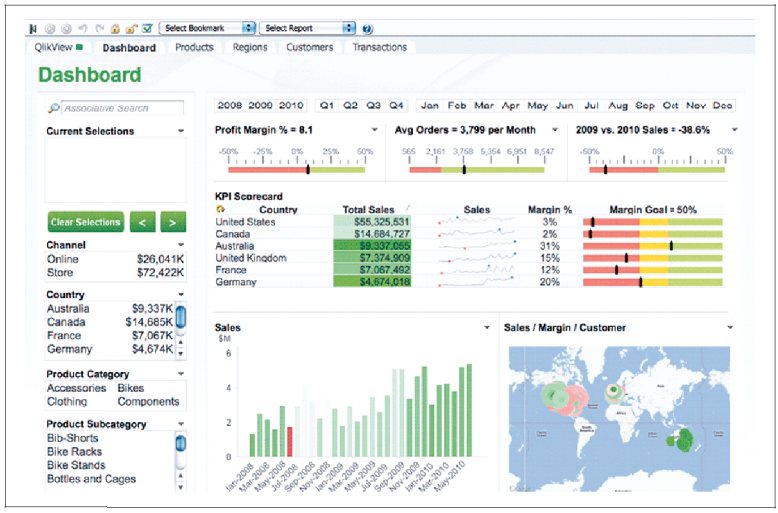


Рис. 7.13. Дашборд для топ-менеджмента компании на платформе QlikView (http://www.qlik.com/us/) показывает KPI по продажам в региональном разрезе

Аналитические дашборды (рис. 7.14) отражают основные тенденции развития и показатели в рамках одного подразделения компании или направления деятельности, например цепочку продаж, маркетинг или цепочку поставок. Обычно они имеют интерактивный характер и дают пользователю возможность тщательного изучения необычного тренда или резко отличающихся показателей, а также позволяют находить данные.

В основном аналитические дашборды используют в своей работе аналитики и руководители подразделений.

Наконец, операционные дашборды (рис. 7.15) дают подробное представление об отдельных аспектах ведения бизнеса, таких как, например, объем продаж в режиме реального времени, интернет-трафик, практические случаи при работе с клиентами или время ожидания, когда вы пытаетесь дозвониться клиенту. Обычно они используются для оповещения, а также в работе сотрудников, которые могут

предпринять немедленные действия, например подключить дополнительные серверы, переключить коллег с выполнения одной задачи на другую, чтобы сократить количество необработанных заказов.

Рис. 7.14. Пример аналитического дашборда о посетителях сайта от Google Analytics

С учетом перечисленных типов дашборды должны использоваться целевым образом. Необходимо четкое понимание, кто ими пользуется и какая информация требуется. Как и в предыдущем разделе, здесь применяется принцип KISS (Keep it simple, Stupid! — Чем проще, тем лучше!)[1]: каждая диаграмма и каждый показатель, которые появляются в дашборде, должны быть обоснованы. Иными словами, не поддавайтесь соблазну добавить туда как можно больше всего. Если дашборд будет перенасыщен данными, интерпретировать эти данные станет сложнее, и он будет менее эффективным. Лучше меньше, да лучше.

[1] URL: https://en.wikipedia.org/wiki/KISS_principle.

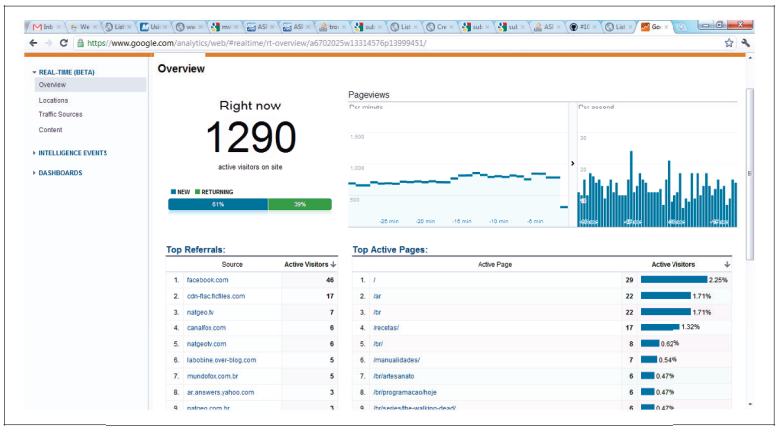


Рис. 7.15. Пример операционного дашборда. Он также сформирован при помощи Google Analytics, но представляет информацию более детально, чем на рис. 7.14. Здесь отражается активность посетителей сайта почти в режиме реального времени: откуда они пришли, на какие страницы направляются, общее число пользователей

Источник: http://www.blog.narensportal.com/2011/12/google-analytics-real-time.html

Ди Джей Патиль и Хилари Мейсон полагают, что имеет смысл использовать несколько дашбордов, отражающих данные в одной области, но для разных категорий пользователей и разных временны́х шкал[1]. Например, в компании One Kings Lane сотрудники службы по работе с клиентами, отвечая на телефонные звонки, могут наблюдать за данными на операционном дашборде, который расположен на настенном мониторе и отражает основные показатели, например число вызовов в режиме реального времени, время ответа и количество решенных проблем клиента. Их руководитель имеет доступ к более детальному аналитическому дашборду, в котором он может систематизировать данные по группе, отдельному заказчику и типу заказа. В дополнение к этому показатели более высокого уровня включены в дашборд для топ-менеджмента, и руководители могут наблюдать за ними в течение дня. В каждом из этих случаев дашборд отвечает целям и задачам тех людей, которые им пользуются.

В контексте этой книги полезно проанализировать, действительно ли дашборды используются для процесса принятия решений. Как уже упоминалось, операционные дашборды отражают изменения (почти)

[1] URL: http://www.oreilly.com/data/free/data-driven.csp.

в режиме реального времени и часто настроены таким образом, чтобы оповещать конечных пользователей о ситуациях, в которых они могут предпринять немедленные действия. Например, если интенсивность телефонных звонков, поступающих в кол-центр компании, увеличивается, руководитель может перенаправить ресурсы из других подразделений, чтобы справиться с наплывом. При этом аналитические и стратегические дашборды практически никогда не бывают единственным источником информации при принятии важных бизнес-решений. Ниже приведены выводы одного из недавних отчетов[1].

> Довольно редко один отчет или дашборд, содержащие аналитическую информацию, служат основой для принятия важного решения. Гораздо чаще пользователи задаются вопросом: почему? Почему в северо-восточном регионе продажи упали на 30%? Почему розничные продажи продукта взлетели в IV квартале? С помощью интерактивных возможностей проведения анализа, которыми располагают опытные пользователи инструментов бизнес-аналитики, можно вовремя задавать эти важные вопросы и так же своевременно получать на них ответы.

Подробнее о процессе принятия решений мы поговорим в главе 9.

Отслеживание использования

Возможно, дашборд бесполезен сам по себе, но он точно будет таковым, если его никто не использует (хотя он может быть бесполезен, и если его используют, но при этом не происходит никаких изменений). В интервью с Кевином Роузом в 2001 году Джек Дорси, сооснователь Twitter и CEO компании Square, высказал интересную мысль:

> У нас в Square есть дашборд и есть показатель «сколько раз сотрудники взглянули на эту панель, чтобы узнать, как обстоят дела в компании». Это говорит о том, насколько сотрудников волнует, как дела у компании[2].

Конечно, компания с управлением на основе данных может пользоваться не только дашбордами. Если отчеты отправляются заинтересованным лицам с сервера, можно настроить показатель, отражающий

[1] URL: http://aberdeen.com/research/9200/RR-holisticBI.aspx/content.aspx.

[2] URL: https://www.youtube.com/watch?v=DQy_HFHOZug.

«уровень открытия» сообщений получателями. Авинаш Кошик идет еще дальше и предлагает «отключать ежеквартально все автоматические отчеты в случайный день/неделю/месяц, чтобы оценить их использование/ценность»[1].

Основные выводы

Мы провели лишь поверхностный обзор сторителлинга и визуализации данных. И вновь я рекомендую обратиться к экспертам. Моя цель была лишь в том, чтобы убедить вас в важности этих вопросов для компании с управлением на основе данных. Проведение аналитической работы и формирование выводов на ее основе — огромный труд. К сожалению, слишком часто кустарно подготовленные презентации не оставляют интересным и важным историям ни малейшего шанса. Навыками визуализации и презентации данных в состоянии овладеть любой, и это станет по-настоящему ценной инвестицией в развитие аналитического направления в компании.

В 1657 году известный французский математик и физик Блез Паскаль в своем сборнике «Письма к провинциалу»[2] отмечал: «Я написал несколько длиннее обычного, потому что у меня не было времени сделать это короче». Его идея, без сомнения, состояла в том, что требуется потратить время и приложить усилия, чтобы отредактировать написанное, выделить основную мысль, убрать все лишнее и оставить только суть. То же самое верно в отношении визуализации данных и сторителлинга.

Стефани Эвергрин выделяет следующие цели презентации данных:

— убедить других;

— оформить мысль;

— стимулировать действие.

Для достижения любой из этих целей необходимо избавиться от всего «графического мусора» и показать пользователю, на чем ему следует сфокусировать внимание. При этом вы не должны заставлять его думать. Важно, что это не означает чрезмерного упрощения содержания.

---

[1] URL: https://www.kaushik.net/avinash/create-analysis-ninjas-data-driven-cultures/.

[2] «Письма к провинциалу» (фр. Lettres Provinciales) — сборник из 18 писем полемического характера Блеза Паскаля, опубликованных в 1656–1657 годах.

Во-первых, начните с четкого понимания вопроса, на который вы пытаетесь ответить, а также с четко сформулированных ожиданий аудитории.

Во-вторых, тщательно подойдите к выбору средств презентации, чтобы они отвечали характеру данных и максимально эффективно могли донести ее посыл.

В-третьих, выделите одно основное сообщение для каждого визуального средства, таблицы или слайда. Предлагайте слушателям информацию по кусочкам, которые они в состоянии «проглотить». Когда де Куббер переделывал слайды программы PRISM, он поместил хронологическую последовательность присоединения разных компаний к программе на одном слайде, а информацию о стоимости программы, которая составила 20 млн долл., — на другом. Таким образом, оба этих информационных блока легко усваиваются. Мне часто приходится сталкиваться с огромными таблицами, содержащими финансовые данные. Обычно они буквально ими набиты: набор финансовых показателей по каждому месяцу за последний год с фактическими параметрами и бюджетами, сравнением месяц к месяцу и год к году и так далее. К сожалению, множество историй, которые могут рассказать эти данные, буквально погребены под грузом самих данных. Возможно, пара ячеек каким-то образом выделены, но приходится просмотреть океан информации, прежде чем добраться до заголовков рядов и столбцов, чтобы получить контекст. Я рекомендовал бы, чтобы аналитик определил историю, которую он хочет донести до остальных, и вынес самую важную информацию — «лакомые кусочки» — на отдельные слайды. Уберите всю «воду» и оставьте только ключевую информацию и ее интерпретацию. Пусть слушатели презентации испытают информационные ощущения, сравнимые с гастрономическим удовольствием от еды из мишленовского ресторана.

В-четвертых, добавьте полезные указатели, такие как название слайда, названия осей, используйте выделение цветом (см. контрольный список, приведенный ранее) для обеспечения нужного контекста. Затем отформатируйте эти указатели так, чтобы они легко воспринимались. Например, не заставляйте зрителей презентации сворачивать шеи, чтобы прочитать вертикально размещенный текст, или напрягать зрение в попытках разглядеть мелкий шрифт.

В-пятых, исключите любые умственные упражнения или вычисления, которые должен произвести слушатель презентации, чтобы связать разрозненные выводы или получить скрытое в данных послание.

Один из примеров — неудобное размещение легенды на столбиковой диаграмме, в результате чего слушатель презентации вынужден, как выразилась Стефани Эвергрин, заниматься «ментальной гимнастикой», чтобы соотнести названия, а следовательно, и смысл столбцов, с их значениями. Еще один пример — сравнение столбцов, на этот раз от Стивена Фью, которое он называет анализом отклонения. Представьте столбиковую диаграмму, которая отображает реальные показатели и запланированные для ряда подразделений компании. Если цель в том, чтобы показать дельту между каждой парой значений, то фактически вы предлагаете слушателям презентации самостоятельно вычислить эту разницу. Подход, который позволит быстрее и легче воспринять эту информацию, заключается в том, чтобы провести все вычисления и представить уже определенные дельты, а не первоначальные пары столбцов. Сконцентрируйтесь на том, что вы хотите показать, что вы хотите, чтобы пользователи вынесли после этой презентации, а затем поставьте себя на их место: что им нужно сделать, чтобы получить это сообщение? Исключите любые задачи, требующие усилий с их стороны.

Если вы выполните все это — проведете зрителя/читателя через один или несколько простых информационных блоков и выводов, — получится более простая и убедительная презентация, способная донести ваше основное сообщение эффективно и без искажения смысла.

Это была заключительная глава из трех, посвященных показателям, типам анализа и презентации результатов, которые составляют суть работы аналитика. В следующей главе мы обсудим важный аспект корпоративной культуры компании с управлением на основе данных — тестирование. То есть мы сосредоточимся на развитии корпоративной культуры под девизом «Докажи это!», в которой идеи тестируются в реальных обстоятельствах на реальных клиентах, и это обеспечивает самые прямые доказательства влияния предложенного изменения или новой характеристики продукта.

ГЛАВА 8

A/B-тестирование

Тот, кто последовательно применяет #abtesting для принятия решений на основе данных, неизменно бывает удручен низким коэффициентом успешности идей.

Рон Кохави

Я усвоил тот факт, что эксперименты, данные и тестирование нужны не для доказательства моей правоты <…> Фактически, чтобы выбрать правильный ответ, мне нужна информация, полученная в результате этого тестирования.

Пи Джей Маккормик[1]

В 1998 году Грегу Линдену, одному из разработчиков Amazon на заре становления этого интернет-гиганта, пришла идея: почему бы не давать пользователям рекомендации при покупке? Супермаркеты раскладывают сладости на полках возле касс, чтобы стимулировать импульсивные покупки, и это работает. Почему бы не заглянуть в корзину пользователя на Amazon.com и не предложить ему персональную рекомендацию, которая может оказаться ему полезна? Линден создал прототип, убедился в его работоспособности и показал всем. О дальнейшем развитии событий лучше услышать из его уст:

> В целом идея была воспринята положительно, но были некоторые затруднения. В частности, старший вице-президент по маркетингу выступал категорически против. Его основное возражение состояло в том, что это может отпугнуть пользователей, которые не захотят оформлять заказ, — это правда, что пользователи часто не завершают процесс покупки онлайн,— и он склонил остальных на свою сторону. На тот момент мне запретили продолжать работу в этом направлении. Мне сказали, что Amazon еще

[1] McCormick PJ. Challenging Data Driven Design, WarmGun 2013, 27 ноября 2013 года. URL: https://www.youtube.com/watch?v=caOIdA9jnQg.

не готова к запуску подобного сервиса. На этом следовало бы остановиться.

Но не тут-то было. Я подготовил сервис для онлайнового тестирования. Я верил в силу рекомендаций, и мне хотелось измерить их влияние на продажи. Говорят, старший вице-президент был в бешенстве, когда узнал, что я готовлю эксперимент. К счастью, даже топ-менеджерам его уровня сложно препятствовать тестированию. Измерения — это всегда хорошо. Единственный весомый аргумент против, что негативный эффект от этого теста мог бы оказаться настолько сильным, что Amazon бы этого не выдержала. Вряд ли такое можно было утверждать, а потому я провел тестирование.

Результат говорил сам за себя. Этот сервис оказался не только востребованным, но разница в уровне продаж была настолько значительной, что отсутствие ее на Amazon в полном масштабе обходилось компании в кругленькую сумму упущенной выгоды. Все заторопились, но теперь уже чтобы запустить рекомендательный сервис для корзины пользователя.

Грегу очень повезло. Даже не в том, что его идея сработала (хотя, разумеется, это важно), а в том, что уже тогда компания Amazon располагала достаточной инфраструктурой для тестирования и такой корпоративной культурой, благодаря которой можно было добиться проведения этого теста. У него получилось доказать ценность своей идеи, реализовать ее на практике и повысить прибыль компании.

Во многих ситуациях, особенно новых для нас, интуиция не всегда срабатывает верно. Часто мы бываем удивлены результатом. Не верите? Тогда возьмем несколько быстрых примеров из онлайн-экспериментов. Первый пример — предложение о покупке в рекламном объявлении. С точки зрения количества переходов (индекс CTR), какое из них сработает лучше и насколько?

— Получите скидку 10 долл. с первой покупки. Заказывайте онлайн сейчас!
— Получите дополнительную скидку 10 долл. Заказывайте онлайн сейчас.

На практике второй вариант оказался эффективнее первого, его индекс CTR был в два раза выше[1]. А как насчет пары объявлений на рис. 8.1?

[1] Gabbert A. The Importance of A/B Testing: 24 Marketing Experts on Their Most Surprising A/B Test, September 25, 2012. URL: http://www.wordstream.com/blog/ws/2012/09/25/a-b-testing.

(Кстати, вы заметили, чем они отличаются?) Какое сработает лучше и насколько?

Вариант слева, грамматически верный благодаря добавлению одной-единственной запятой, был на 8% эффективнее.

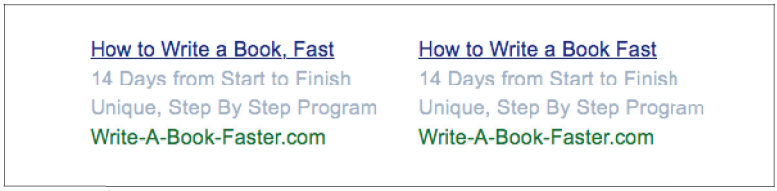


Рис. 8.1. У какого из этих вариантов индекс CTR будет выше? У грамматически правильного объявления слева индекс CTR на 8% выше (4,4% по сравнению с 4,12%).

Наконец, в заключительном примере (рис. 8.2) даны две практически идентичные версии интернет-страницы — за исключением того, что в варианте слева все поля в форме для заполнения необязательные. У этого варианта коэффициент конверсии был на 31% выше. Более того, качество этих контактов было *выше*.

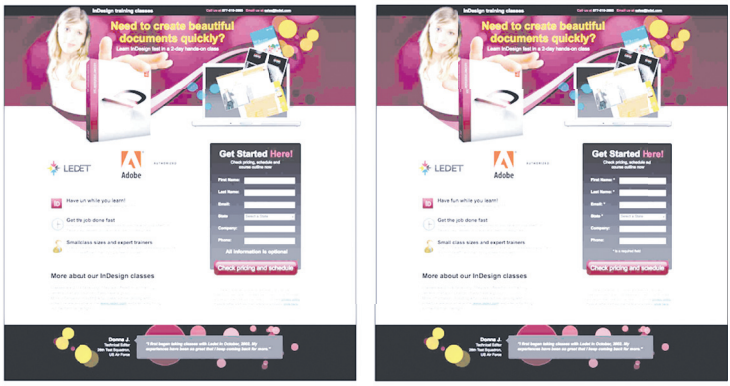


Рис. 8.2. В варианте слева все поля формы для заполнения необязательны. Коэффициент конверсии этого объявления на 31% выше, более того, качество этих контактов тоже было выше

Источник: https://www.behave.org/test/adobe-training-company-tests-required-form-fields-vs-not-required-%E2%80%93-which-version-got-a-31-lift-in-lead-gen-form-submissions/

Во всех этих примерах было сложно прогнозировать, какой вариант окажется эффективнее, и еще сложнее было предсказать влияние на другие показатели. Именно поэтому качественно подготовленный эксперимент имеет такую ценность. Он переводит диалог из плоскости «Мне кажется…» в плоскость «Согласно данным…». Таким образом, это неоценимый компонент компании с управлением на основе данных.

Давайте рассмотрим этот аспект в перспективе. В главе 5 мы провели обзор пяти видов анализа, включая каузальный анализ, являющийся вершиной аналитической работы, по крайней мере, с точки зрения обычного бизнеса. Контролируемый эксперимент, применение научного метода или «научных методов работы с данными»[1] — прямой способ выявить эти причинно-следственные отношения.

Три примера, обсуждавшихся выше, представляли собой варианты эксперимента под названием A/B-тестирование. Сейчас я приведу краткое его описание. Какие-то подробности и детали я добавлю чуть ниже в этой главе, а сейчас опишу основную идею. При проведении A/B-тестирования вы устанавливаете контроль, например, над текущим состоянием сайта (вариант A). Половину трафика своего сайта вы направляете на эту версию. Эти посетители сайта будут относиться к группе A. Вторую половину пользователей вы направляете на другую версию сайта, имеющую небольшие отличия, например, надпись на кнопке для оформления заказа — «Приобрести», а не «Купить сейчас» (вариант B). Эти посетители сайта относятся к группе B. Вы определяете показатель, который хотите протестировать, например влияет ли надпись на кнопке на уровень средней выручки на посетителя. Вы проводите эксперимент в течение установленного времени (дней или недель), а затем осуществляете статистический анализ. Вы анализируете, отмечается ли статистически значимая разница в фокусном поведении — в данном случае в показателе выручки на посетителя — между группой A и группой B. Если разница есть, то в чем ее причина? Если эксперимент был полностью контролируемым (то есть в условиях имелось лишь одно небольшое отличие), возможны два варианта. Это могла быть случайность, что вероятно при слишком маленьком размере выборки (то есть эксперимент не соответствовал стандартам). Или же разница между вариантами A и B носит причинно-следственный характер. Согласно данным, фактор, который отличался, вызвал изменение поведения пользователей.

Поскольку объективное проведение экспериментов и их влияние на корпоративную культуру — критически важный фактор для компании

[1] Patil D. J. and Mason H. Data Driven: Creating a Data Culture. Sebastopol, CA: O'Reilly, 2015.

с управлением на основе данных, эта глава будет посвящена A/B-тестированию. Мы охватим оба подхода: более распространенный классический частотный подход, а также более современный байесовский подход. Мы подробно разберем, как проводить тесты, на примерах того, как это делать и как этого делать не стоит. Помимо примеров, описанных ранее, я приведу еще ряд примеров, позволяющих понять, зачем нам все это нужно и какое существенное влияние это может оказать на бизнес. Итак, приступим.

Почему A/B-тестирование?

Как уже говорилось, наша интуиция может нас подвести (подробнее к этому мы еще вернемся в главе 9). Даже эксперты в определенных областях ошибаются чаще, чем им бы хотелось это признать. В своей книге A/B Testing: The Most Powerful Way To Turn Clicks Into Customers (Wiley & Sons) Дэн Сирокер, генеральный директор и создатель платформы для A/B-тестирования Optimizely, рассказывает о некоторых аспектах работы своей компании в 2008 году во время предвыборной кампании Барака Обамы. Перед ними стояла задача оптимизировать интернет-страницу для потенциальных сторонников Обамы и с ее помощью собрать базу адресов электронной почты этих людей. Изначально на странице была размещена статичная картинка с красной кнопкой с надписью «SIGN UP» («ПОДПИСАТЬСЯ»). Команда разработчиков полагала, что видеоролики с самыми убедительными выступлениями будут привлекать пользователей эффективнее статичного изображения. После того как были протестированы разные статичные картинки и разные видеоролики, стало ясно, что «*любой* видеоролик значительно уступает *любому* изображению». Оптимальное сочетание изображения и надписи на кнопке (лучшим вариантом оказался «LEARN MORE» («ПОДРОБНЕЕ») повысило уровень подписки на 40,6%. Это соответствовало дополнительно почти 2,8 млн подписчиков, 280 тыс. волонтеров и невероятным 57 млн долл. дополнительных пожертвований. Бывает невозможно предугадать, что и как именно сработает: поведение людей непостоянно и непредсказуемо. Тем не менее результаты, подобные этим, показывают, что мы можем получить важное конкурентное преимущество и непосредственно узнать своих текущих и потенциальных клиентов.

Более того, онлайн-тестирование — относительно недорогое и простое. Не обязательно требуются новые технологии и творческие усилия, чтобы сделать новую версию надписи на кнопке «ПОДРОБНЕЕ»

вместо «ПОДПИСАТЬСЯ». Кроме того, эти изменения не навсегда. Если вы что-то попробовали, но это не сработало, просто вернитесь к первоначальному варианту. В любом случае вы узнаете что-то новое о своих клиентах. Вы практически ничем не рискуете.

Предметом тестирования может стать все что угодно. В какой бы отрасли вы ни работали, всегда есть что оптимизировать и имеются уроки, которые можно извлечь. Команда, работавшая на предвыборный штаб Обамы, проводила множество самых разных тестов. Она тестировала темы сообщений в электронных рассылках, содержание рассылок, время отправления и частоту, все аспекты сайта, даже сценарии, на которые волонтеры опирались в беседе с потенциальными донорами. Как показывает этот пример, подобное тестирование может не ограничиваться только онлайн-форматом. В качестве еще одного примера можно привести маркетинговые акции по увеличению лояльности покупателей, когда компания неожиданно дарит подарки определенной категории клиентов. Эти акции следует тщательно продумывать. С их помощью можно сравнивать такие показатели, как процент возврата, «пожизненная ценность клиента», а также положительные отзывы в социальных сетях от тех, кто получил подарок, и тех, кто не получил. Во всех этих случаях к экспериментам следует относиться с таким же уровнем научной строгости и структурировать их с той же тщательностью, что и онлайн A/B-эксперименты.

Один из приятных аспектов A/B-тестирования в том, что вам не требуется предварительного причинно-следственного объяснения, почему что-то должно сработать. Нужно просто провести тест, изучить результаты и найти те факторы, которые обеспечивают позитивное влияние. Кохави отмечает, что в Amazon половина экспериментов не приносила результатов, а в Microsoft — две трети[1]. Чтобы выиграть в долгосрочной перспективе, совсем не обязательно, чтобы срабатывал каждый эксперимент. Единственное положительное изменение способно оказать огромное влияние на итоги всей деятельности.

Практические рекомендации по A/B-тестированию

После такого вступления, описавшего преимущества применения A/B-тестирования, давайте перейдем к практическим аспектам и посмотрим, как качественно его организовать.

[1] Kohavi R. Planning, Running, and Analyzing Controlled Experiments on the Web, June 2012. URL: http://bit.ly/kohavi-planning.

ПОДГОТОВИТЕЛЬНЫЙ ЭТАП

В этом разделе мы рассмотрим ряд аспектов, на которые следует обратить внимание в ходе подготовительного этапа. Первое и самое важное — сформулировать критерии, которыми вы будете руководствоваться. Затем мы рассмотрим так называемые A/A-тесты, которые важны для проверки аппарата эксперимента. Кроме того, их можно использовать для генерирования нескольких ложноположительных результатов, чтобы наглядно продемонстрировать руководителям и коллегам статистическую значимость и важность достаточно большой выборки. Далее мы детально изучим план A/B-теста (что мы тестируем, кто участники, какой анализ будет проводиться и так далее). Наконец, мы остановимся на важнейшем аспекте и фактически первом вопросе, который задают все новички: каким должен быть размер выборки?

Критерии эффективности

Рекомендация: четко сформулируйте критерии эффективности до начала тестирования.

Важно иметь четкое понимание своей цели и имеющихся средств. Зачем мы это делаем? Особенно важно до начала тестирования определить ключевые показатели, которые иногда называют критериями общей оценки. В чем будет заключаться успешный результат? Если вы этого не сделаете, у вас может появиться соблазн собрать как можно больше данных в ходе эксперимента, а на этапе анализа начать статистически тестировать всё и ухватиться за значимые результаты. Хуже того, может появиться мысль выборочно отразить в отчетах только положительные показатели и результаты. Такой подход лишь доставит вам неприятности и не принесет долгосрочной пользы компании.

A/A-тестирование

Рекомендация по проведению A/A тестов

Если A обозначает контрольную группу, то, как вы уже могли догадаться, A/A-тестирование представляет собой сравнение двух контрольных групп, все изначальные условия для которых одинаковые. Какой в этом смысл? На самом деле есть целый ряд преимуществ.

Во-первых, вы можете применять его для тестирования и мониторинга вашей инфраструктуры и процессов распределения. Если вы зададите

настройки системы для разделения трафика 50/50, но размер выборок в двух группах будет сильно отличаться, это означает, что с вашим процессом распределения что-то не так.

Во-вторых, если при сопоставимом размере двух выборок наблюдаются сильно отличающиеся показатели деятельности, это свидетельствует о проблеме с отслеживанием событий, проблеме при проведении анализа или составлении отчетности. При этом можно ожидать уровень различий при A/A-тестировании около 5%, сделав допущение, что вы придерживаетесь стандартного статистического уровня значимости 5%. Что действительно нужно отслеживать при многократном проведении A/A-тестов, так это наблюдаются ли у вас значительные расхождения, на порядок больше, чем стандартный уровень значимости. Если да, это может свидетельствовать о проблеме. Однако Георгий Георгиев резонно отмечает: «Даже если вам требуется всего 500 или 100 A/A-тестов, чтобы заметить статистически значимые отклонения от ожидаемых результатов, это все равно огромная потеря денег. Просто потому, что впечатления, клики, посетители — это все не бесплатно, не говоря уже о том, как вы могли бы использовать этот трафик»[1]. Нужно проводить множество A/B-тестов и постоянно внедрять инновационные решения. Однако, если у вас нет постоянного потока A/B-тестов или возник перерыв, проводите A/A-тесты.

В-третьих, результаты тестирования можно использовать для оценки вариативности тех показателей, которые вы контролируете. В некоторых вычислениях размера выборки, таких как при тестировании среднего значения (скажем, средний размер корзины или время, проведенное на сайте), это значение понадобится для вычисления размера выборки.

Наконец, в блоге Nelio A/B Testing отмечается, что применение A/A-тестов имеет, помимо прочего, и образовательную функцию[2]. Для тех компаний, где конечные пользователи или руководители никогда раньше не имели дела с A/B-тестированием и не особо подкованы в вопросах вероятности и теории статистики, это будет весьма полезно. Не стоит торопить события и сразу переходить к A/B-тестированию, полагая, что тестируемые показатели должны быть лучше контрольных, даже когда результаты впечатляют. Статистически значимый результат может быть делом случая, и самое наглядное доказательство этого — A/A-тестирование.

---

[1] URL: http://blog.analytics-toolkit.com/2014/aa-aab-aabb-tests-cro/.

[2] URL: https://neliosoftware.com/blog/the-importance-of-aa-testing-no-not-a-typo/.

Планирование A/B-теста

Рекомендация: продумайте весь ход эксперимента до его начала.

При планировании теста следует обратить внимание на многие аспекты. Тем компаниям, которые намерены внедрить у себя культуру A/B-тестирования, я рекомендовал бы заранее продумать приведенный ниже спектр вопросов. После того как вы запустите тестирование, обсуждать критерии эффективности будет поздно. Вряд ли вы захотите, чтобы кто-то подтасовывал результаты во время анализа. Этап обсуждения и всех согласований должен предшествовать этапу самого тестирования.

Цель

— В чем цель этого теста?

Зоны ответственности

— Кто представитель от бизнеса?

— Кто отвечает за реализацию тестов?

— Кто осуществляет бизнес-аналитику?

Планирование эксперимента

— Какие показатели вы планируете тестировать, а какие будут являться контрольными?

— Кто составит вашу тестовую и контрольную группы (то есть люди)?

— Каковы ваша нулевая и альтернативная гипотезы?[1]

— Какие показатели вы планируете отслеживать?

— Когда будут обсуждаться результаты и формироваться обратная связь?

— Когда начнется тестирование?

— Требуется ли время для «разогрева»? В таком случае, с какого момента пойдет отсчет эксперимента для аналитических целей?

— Сколько продлится тест?

— Как определили размер выборки?

[1] Нулевая гипотеза — основное предположение об отсутствии разницы между сравниваемыми вариантами (например, CTR в контрольной группе = CTR в тестируемой группе). Альтернативная гипотеза — то предположение, к которому вы придете, если опровергнете нулевую гипотезу. Оно может быть одним из трех типов: CTR контрольной группы != CTR тестируемой группы; CTR контрольной группы > CTR тестируемой группы или CTR контрольной группы < CTR тестируемой группы. Стоит придерживаться двусторонней альтернативной гипотезы (то есть !=), если у вас нет веской причины остановиться на прямой альтернативе (то есть > или <).

Процесс анализа

— Кто будет проводить анализ? (В идеале должно быть разделение между теми, кто планирует эксперимент, и теми, кто оценивает результаты.)

— Какой вид анализа будет проводиться?

— Когда начнется процесс анализа?

— Когда он завершится?

— Какое программное обеспечение будет использоваться для его проведения?

Результаты

— Как будут распространяться результаты анализа?

— Как будет приниматься окончательное решение?

Список кажется довольно длинным, но по мере того как вы будете проводить все больше и больше тестов, некоторые из вопросов и ответов перейдут в разряд стандартных. Например, ответы могут быть: «При проведении анализа мы всегда используем R» или «Проведение статистического анализа входит в обязанности Сары». Этот набор вопросов станет постепенно внедряться в корпоративную культуру, процесс будет становиться все более автоматическим, пока наконец он не станет естественным и привычным.

По получившемуся у меня описанию процедура проведения эксперимента и процесс анализа — очень четкие, почти клинические и доведенные до автоматизма: тест А против теста В, какой тест выигрывает, тот и внедряется на практике. Если бы так и было, то это был бы полный процесс управления на основе данных. Но реальный мир гораздо сложнее. В игру вступают другие факторы. Во-первых, результаты не всегда четко определены. Возможна двусмысленность. Не исключено, что показатель в тестовой группе был немного завышенным на протяжении всего теста, но незначительно. Или некоторые факторы компенсировали друг друга (например, объем продаж и уровень конверсии). Или, возможно, в процессе анализа вы обнаружили фактор, способный повлиять на объективность результатов. Все это может негативно сказаться на их анализе и интерпретации. Подобная двусмысленность вполне реальна. Во-вторых, отдельный эксперимент не обязательно отражает ту долгосрочную стратегию, которой следует компания. Пи Джей Маккормик приводит пример подобной ситуации на Amazon[1]. Он описы-

[1] URL: https://www.youtube.com/watch?v=caOIdA9jnQg.

вает A/B-тест, в котором в качестве контрольного элемента выступало крошечное изображение покупаемого продукта, настолько маленькое, что его было невозможно рассмотреть. В качестве тестируемого элемента было более крупное изображение продукта. Казалось бы, результат теста очевиден. Но не все так просто: маленькое изображение, по которому даже не было понятно, на что кликает пользователь, победило! Тем не менее в компании приняли решение перейти на размер изображения крупнее. Почему?

«Мы запустили более крупные изображения, потому что так пользователи видят, что они покупают. Это более положительный опыт. Кроме того, это совпадает с тем, к чему мы стремимся в долгосрочной перспективе, и с нашим ви́дением. Данные не мыслят в долгосрочной перспективе за вас. Они не принимают решения. Они лишь дают информацию — пищу для размышлений. Но если вы принимаете решения автоматически, не задумываясь о том, что означают эти данные, и не соотнося их с вашим долгосрочным видением относительно вашего продукта или пользователей, то, скорее всего, ваши решения будут ошибочными»[1].

(Процесс принятия решений будет темой следующей главы.)

Размер выборки

Рекомендация: используйте калькулятор размера выборки.

Вопрос, который мне чаще всего задают относительно A/B-тестирования: «Как долго нужно проводить тестирование?» Обычно я отвечаю: «Я не знаю, нужно подсчитать с помощью калькулятора размера выборки».

Этот раздел более технический по сравнению с остальными, а потому те, кого статистика приводит в ужас, могут просто его пропустить. Основной вывод в том, что вам необходимо рассчитать минимальный размер выборки с помощью простого статистического онлайн-инструмента и придерживаться этого размера. Нельзя досрочно прекратить тестирование и рассчитывать на значимые результаты.

[1] Это делает обоснованным вопрос: зачем вообще проводить тестирование? Если результаты тестирования не стимулируют действий, насколько это рациональная трата времени и сил?

Причина, по которой непросто дать ответ на этот вопрос, заключается в том, что существует множество факторов, которые мы пытаемся оптимизировать.

Предположим, мы проводим стандартный A/B-тест. Есть четыре возможных сценария. Между сравниваемыми показателями не наблюдается различия, тогда:

1) мы приходим к *верному* заключению, что различия нет;
2) мы приходим к *ошибочному* заключению, что различия нет; это ложноположительный результат.

Или между сравниваемыми показателями наблюдается различие, тогда:

3) мы приходим к ошибочному заключению, что различия нет; это ложноотрицательный результат;
4) мы приходим к верному заключению, что различие есть.

Вышесказанное можно суммировать следующим образом.

Истина			
Ваши результаты	*Нет различия*	*Нет различия* 1. Верный результат	*Есть различие* 3. Ложноотрицательный результат
	Есть различие	2. Ложноположительный результат	4. Верный результат

Наша цель — попытаться оптимизировать вероятность верного заключения (1 или 4) и минимизировать вероятность сделать ложноположительное (2) или ложноотрицательное (3) заключение.

Для этого в нашем распоряжении два рычага, которыми мы можем воспользоваться.

Первый — более очевидный размер выборки. Если бы вы проводили опросы избирателей на президентских выборах, то были бы более уверены в своем прогнозе, если бы опросили 500 тыс. проголосовавших, а не 5 тыс. Это верно и относительно A/B-тестирования. Более значительная выборка повышает вашу статистическую мощность (статистический термин) при определении статистически достоверного различия, если это различие действительно существует. Возвращаясь к нашему примеру с четырьмя возможностями, если различие есть, то более крупная выборка снижает вероятность ложноотрицательного заключения (то есть более вероятно сделать вывод 4, чем 3). Обычно

используется мощность 0,8. Это означает, что при существовании различия мы сможем определить его с вероятностью 80%. Запомните это, мы вернемся к этому чуть позже.

Второй рычаг в нашем распоряжении — это статистический уровень значимости, обычно составляющий 5%[1]. (Для масштабной выборки хороший подход — выбрать $p \leq 10^{-4}$.) Это означает приемлемую вероятность сделать ложноположительное заключение, если на самом деле различия между сравниваемыми показателями нет. Предположим, у нас есть обычная монета. Мы подбросили ее десять раз, и десять раз выпал орел. Кажется, сюда закралась погрешность в пользу орла. Но самая обычная монета все же могла бы упасть орлом вверх десять раз подряд, но только один раз из 1024 раз, или примерно 0,1% от всех случаев. Если мы предположим, что монета с погрешностью, то рискуем ошибиться в 0,1% случаев. Это кажется приемлемым риском. Далее, предположим, мы решаем, что если мы увидим восемь, девять или десять орлов или, наоборот, ноль, один или два орла, то сделаем вывод, что монета с погрешностью. При этом есть вероятность ошибиться уже в 11% случаев. Это кажется слишком рискованным. Суть в том, чтобы сбалансировать убедительность доказательства, что тестируемое качество действительно оказывает влияние, против вероятности, что мы наблюдаем лишь случайный эффект (а фактического различия нет).

Итак, вооружившись критерием статистической мощности = 0,8 и уровнем статистической значимости = 5%, переходим к калькулятору размера выборки (рис. 8.3). Вводим два этих значения (см. нижнюю часть рисунка), но кроме этого нужно предоставить дополнительную информацию. Этот тип калькулятора (оптимизированный для определения конверсии, то есть контроля перехода на сайт) запрашивает базовый показатель коэффициента конверсии. Это значит текущий коэффициент в вашей контрольной группе. Он также запрашивает значение минимального заметного эффекта. Это означает, что при существовании значительного различия, например 7%, вы сможете определить его сразу же и обойтись при этом небольшим размером выборки. Если требуется определить менее значительное различие, например 1%, потребуется выборка более крупного размера, чтобы убедиться, что различие действительно существует и оно не случайно.

[1] Почему 5%? Чаще всего его связывают с единственным предложением из работы Р. Фишера 1925 года, но на самом деле история начинается в 1881 году с Ф. Бесселя. Эта история описана в моем блоге. URL: http://www.p-value.info/2013/01/whats-significance-of-005-significance_6.html.

При коэффициенте конверсии 10% и различии 1% вам потребуется выборка из 28 616 человек: 14 313 составят контрольную группу и столько же — тестовую.

Есть разные калькуляторы размера выборки, подходящие для разных ситуаций. Например, для сравнения средних значений, скажем, среднего размера корзины в контрольной группе и тестовой группе, калькулятор размера выборки будет похожим, но требования по вводимой информации станут слегка отличаться, например базовым показателем вариативности[1].

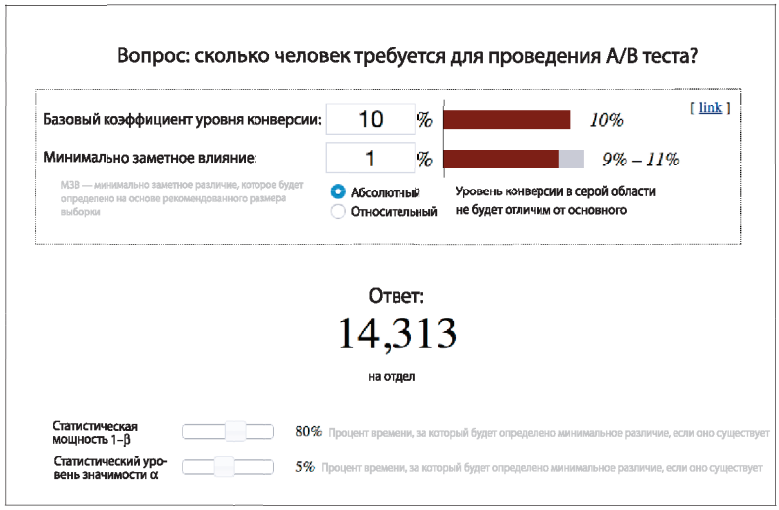


Рис. 8.3. Калькулятор размера выборки для определения конверсии

Источник: http://www.evanmiller.org/ab-testing/sample-size.html

Оценить, сколько дней нужно на проведение эксперимента, можно путем деления среднего дневного трафика на общий размер выборки.

Обратите внимание, что это *минимальный* размер выборки. Предположим, исходя из размера выборки и уровня посещаемости вашего сайта, вам рекомендуется проводить тестирование в течение четырех дней. Если в эти дни уровень посещаемости сайта был ниже обычного среднего показателя, следует продолжить эксперимент, пока вы не достигнете минимального размера выборки. Если вы не продлите эксперимент или слишком рано его завершите, результаты будут необъективными. В итоге у вас повысится вероятность получить ложноотрицательное

[1] URL: http://www.biostat.ucsf.edu/sampsize.html.

заключение: вы не сможете определить различие, которое существует. Более того, если наблюдается положительный результат, повышается вероятность того, что он не отражает действительность (см. Most Winning A/B Test Results Are Illusory[1]). Это чрезвычайно важный эффект. Вы видите положительное влияние, празднуете свою победу, запускаете тестируемую характеристику в массовое производство, а затем не наблюдаете никакого роста. Итог — напрасно потраченные время и силы, а кроме того, утрата доверия.

Итак, мы определили размер выборки и продолжительность тестирования. Или не совсем? Если вы проводите тестирование в течение четырех дней с понедельника по четверг, получите ли вы те же самые демографические и поведенческие характеристики пользователей, которые получили бы, проводи вы тестирование с пятницы по понедельник? В большинстве случаев они будут различаться. Это «эффект дня недели» в действии: пользователи, посещающие сайт в выходные, и их поведение отличаются от тех, что посещают сайт в другие дни. Таким образом, если согласно калькулятору размера выборки тестирование рекомендуется проводить в течение четырех дней, лучше продлите его еще на три дня, чтобы охватить неделю полностью. Если рекомендуемая продолжительность тестирования — 25 дней, проводите его в течение четырех недель.

Как видите, определение размера выборки — важный аспект. Если вы захотите обойтись выборкой меньшего размера, чем необходимо, то, скорее всего, получите ложные результаты: они будут указывать на наличие положительного эффекта, но не смогут генерировать дополнительную прибыль. Или, наоборот, вам не удастся определить наличие эффекта от тестируемой характеристики и вы столкнетесь с упущенной выгодой. Очевидно, оба этих варианта развития ситуации нежелательны. Наконец, расчеты размера выборки иногда бывают сложными, и для качественной оценки без калькулятора не обойтись. Воспользуйтесь имеющимися у вас инструментами.

ПРОВЕДЕНИЕ ТЕСТИРОВАНИЯ

После того как вы определили тестируемую характеристику и настроили на сайте инструменты для сбора необходимых данных, переходим к следующим вопросам: кто будет участвовать в тестировании, когда оно начнется и когда завершится?

[1] URL: http://www.qubit.com/sites/default/files/pdf/mostwinningabtestresultsareillusory_0.pdf.

Выбор участников тестирования

Рекомендация: предложите оценить тестируемую характеристику 50% пользователей, отвечающих критериям отбора, и обеспечьте стабильность процесса.

Первый вопрос, возникающий при выборе участников тестирования, — это критерии отбора. Возможно, некоторые пользователи не должны принимать участие в тестировании вообще. Во многих случаях при проведении A/B-тестирования ориентируются на всех посетителей сайта. Но вполне возможно, что вас интересует только конкретная категория посетителей, например только те, кто совершает повторные покупки, или пользователи из конкретного региона или с определенными демографическими характеристиками. Все зависит от тестируемой характеристики и целевой аудитории. Критерии отбора должны быть четко определены.

Эта выборка пользователей представляет совокупность всех участников тестирования, которых можно разделить на две группы — контрольную и тестовую. Следующий вопрос: в каком соотношении формировать группы? В идеале совокупный трафик следует разделить 50/50, но так получается не всегда. Кохави и др. отмечают, что «распространенная практика среди новичков, которые только начинают проводить подобные эксперименты, — предложить протестировать новую характеристику лишь небольшому проценту пользователей»[1]. Вероятно, они поступают так, чтобы избежать риска и снизить негативное влияние, если с новой характеристикой возникнут проблемы. Однако это плохая стратегия, так как тогда проведение тестирования займет больше времени. Тестирование должно пройти для минимального размера выборки для обеих групп — контрольной и тестовой, поэтому, если трафик в тестовой группе снижен, например, до 10%, очевидно, что потребуется гораздо больше времени, пока размер выборки тестовой группы достигнет требуемого. В этом случае рекомендуется, наоборот, «усилить» эксперимент, повысив пропорцию трафика в тестовой группе (подробнее мы коснемся этого чуть позже), чтобы снизить риск, но достигнуть трафика в 50%.

Необходим надежный механизм распределения посетителей сайта в контрольную или тестовую группу. То есть необходимо сделать это случайным образом, но системно. При рекомендованном делении 50/50 у пользователя должна быть одинаковая вероятность оказаться в любой из двух групп. Один из подходов заключается в применении

[1] http://www.exp-platform.com/documents/controlledexperimentdmkd.pdf.

генератора случайных чисел, назначении пользователям их группы и сохранении этого варианта в определенной базе данных или, возможно, в куки-файле. На основании этой информации пользовательский интерфейс (UI) в дальнейшем будет отображать тот вариант сайта, который нужно для этой группы. Этот подход хорошо работает для сайтов, где все пользователи аутентифицированы. Другой подход состоит в спонтанном распределении пользователей по двум группам. При этом важно, чтобы при повторном возвращении на сайт пользователь системно попадал в одну и ту же группу, поэтому здесь необходим четко определенный процесс распределения пользователей. Например, можно применить мод или подходящую функцию хеширования (расстановки ключей) к каждому ID пользователя. (Кохави и др. подробно обсуждают разные протоколы для системного распределения.) Обеспечение стабильного опыта для пользователя имеет важное значение. Если он будет видеть разные версии сайта, это может привести его в замешательство и повлиять на качество данных и их анализа.

Впрочем, некоторое замешательство может возникнуть в любом случае. Представьте *постоянного* пользователя, который попал в тестовую группу и в первый раз увидел модифицированную версию сайта. У него есть определенные ожидания, сформировавшиеся после предыдущего посещения сайта, и, чтобы осмыслить новый опыт, ему потребуется какое-то время. У пользователя, который посещает сайт в первый раз, еще нет сформированных ожиданий, поэтому ему может быть легче сразу во всем разобраться. Так называемый эффект первичности может быть довольно значительным, и его следует учитывать при проведении анализа данных.

Начало тестирования

Рекомендация: постепенно наращивайте количество пользователей в тестовой группе до 50% от совокупной выборки.

В начале эксперимента можете сразу направить 50% трафика в тестовую группу. Сложность заключается в том, что, если закралась ошибка, в результате которой половина ваших пользователей получила негативный опыт, то вы можете просто потерять эту половину пользователей. Вместо этого можно попробовать другой подход: постепенно наращивать количество пользователей в тестовой группе и тщательно контролировать показатели. Рон Кохави предлагает следующую схему[1]:

[1] URL: http://www.exp-platform.com/documents/controlledexperimentdmkd.pdf.

— 1% пользователей направляется в тестовую группу на четыре часа;
— 5% пользователей направляются в тестовую группу на четыре часа (то есть перевод дополнительных 4% пользователей из контрольной группы в тестовую);
— 20% пользователей направляются в тестовую группу на четыре часа;
— 50% пользователей направляются в тестовую группу на все оставшееся время тестирования.

Конечно, если вы видите, что возникла проблема, у вас должна быть возможность немедленно прекратить тестирование и вернуть весь трафик в контрольную группу.

Завершение тестирования

Рекомендация: проводите эксперимент, пока не охватите минимальный размер выборки или больше.

Я уделил пристальное внимание вопросу определения размера выборки, потому что у него могут быть серьезные последствия. Если завершить тестирование раньше срока, вероятность ошибки существенно возрастет. Можно не увидеть положительного эффекта от тестируемой характеристики, которая могла бы принести компании дополнительную прибыль, или, наоборот, можно приписать случайный положительный опыт эффекту от тестируемой характеристики (то есть имеется риск запустить в массовое производство характеристику, не имеющую никакого эффекта). Иными словами, возрастает вероятность получения ложноположительного или ложноотрицательного результата. Никогда не прекращайте эксперимент досрочно только потому, что наблюдается положительный эффект от тестируемой характеристики.

К сожалению, многие производители программного обеспечения для A/B-тестирования побуждают пользователей проводить эксперимент только до того момента, когда будут достигнуты значимые результаты. *Никогда не проводите тестирование подобным образом!* (Кажется, так я достаточно дал понять, что это действительно важно?) После изучения предложений четырех производителей специализированного ПО Мартин Гудсон отмечает: «Некоторое ПО для A/B-тестирования разработано таким образом, что оно постоянно отслеживает результаты и останавливает процесс, как только достигаются значимые результаты. Однако когда тестирование проводится подобным образом,

вероятность ложноположительного результата может достигать 80%»[1]. (См. также How Not To Run An A/B Test[2]).

Когда вы запустили эксперимент и убедились в отсутствии грубых ошибок, самым разумным будет поступить как Ронко: «Наладьте процесс и забудьте о нем». В период тестирования отслеживайте размер выборки, а не значения показателей.

Другие подходы

Далее я сделаю краткий обзор двух других подходов, которые можно использовать в дополнение к простому A/B- или A/A-тестированию или вместо них.

МНОГОВАРИАНТНОЕ ТЕСТИРОВАНИЕ

До того мы обсуждали только варианты тестирования с двумя переменными: контрольно-тестовое (A/B) или контрольно-контрольное (A/A). Такое тестирование обычно бывает простым и эффективным. Однако у него есть свои недостатки. Вспомните пример с избирательной кампанией Обамы, когда аналитики тестировали разные надписи на кнопке и разные изображения. У них было пять разных вариантов надписи и по крайней мере шесть разных изображений, то есть общее количество разных комбинаций было не меньше 30. Последовательное тестирование всех этих комбинаций заняло бы в 30 раз больше времени, чем проведение одного A/B-теста. Именно по этой причине в некоторых случаях используются многовариантные тесты.

Это тестирование также иногда называют факторным экспериментом, и в ходе него все возможные комбинации тестируются одновременно. То есть группа 1 видит изображение 1 и текст 1, группа 2 — изображение 2 и текст 2, и так до группы 30, которая видит изображение 6 и текст 5.

Какие у этого подхода плюсы и минусы? Если у вас высокая посещаемость сайта и вы можете позволить разделить трафик между разными комбинациями, у вас есть возможность провести тестирование параллельно, то есть потратить на него меньше времени. (Сервис YouTube, у которого, очевидно, огромная посещаемость, в 2009 году провел эксперимент, включавший тестирование 1024 комбинаций[3]. Оптимальная

[1] URL: http://www.qubit.com/sites/default/files/pdf/mostwinningabtestresultsareillusory_0.pdf.

[2] URL: http://www.evanmiller.org/how-not-to-run-an-ab-test.html.

[3] URL: https://youtube.googleblog.com/2009/08/look-inside-1024-recipe-multivariate.html.

комбинация привела к росту количества подписчиков сервиса на 15%.) Кроме того, вы можете протестировать так называемый эффект взаимодействия. Возможно, более крупная кнопка подписки эффективнее, чем кнопка стандартного размера, и кнопка подписки красного цвета эффективнее, чем кнопка подписки синего цвета, а если объединить эти две характеристики, то выяснится, что крупная красная кнопка подписки еще эффективнее, чем просто крупная или просто красная кнопка.

Не все комбинации имеет смысл тестировать. Предположим, первый фактор, который нужно протестировать, — цвет кнопки подписки: красный (текущий) или черный (тестовый). При этом второй фактор — цвет надписи на кнопке: черный (текущий) или белый (тестовый). Общее количество возможных комбинаций — четыре, но комбинация «черная кнопка / черный цвет надписи» явно в тестировании не нуждается. Или, как отмечают Кохави и др., более крупное изображение товара и его дополнительное описание может стать не самой удачной комбинацией, поскольку тогда кнопка для оформления заказа слишком сильно сместится вниз. Подобные моменты нужно отслеживать еще на стадии планирования эксперимента и не включать в тестирование.

Однако даже когда все сформировавшиеся комбинации имеют смысл, вполне возможно провести тестирование на основе выборки из этих комбинаций. Это так называемый дробный факторный эксперимент. Он проводится на основе тщательно сделанной выборки комбинаций, которая позволяет рационально оценить как основной эффект, так и эффект взаимодействия. При этом такой эксперимент сложнее разработать, и он не обеспечивает того уровня информации, которого можно достигнуть с помощью полного многовариантного тестирования или последовательной серии A/B-тестов. Если вы все-таки проводите многовариантные тесты, с их помощью лучше изучать больше факторов (то есть разные типы тестируемых характеристик, таких как изображения и текстовые надписи), чем уровни (то есть разные варианты внутри одного фактора, например пять разных вариантов текста надписи). Кроме того, вам придется играть «по-крупному» и провести тест для 100% пользователей, чтобы максимально увеличить размер выборки и статистическую мощность.

Неудивительно, что анализировать результаты многовариантного теста сложнее: требуется применение более продвинутых статистических инструментов (таких как дисперсионный анализ, или ANOVA), чем те, что используются для проведения A/B-тестирования. Кроме того, визуализировать результаты анализа тоже сложнее.

Итак, многовариантное тестирование позволяет быстрее изучить «пространство проектных решений» или другие аспекты бизнеса, а также проверить эффект взаимодействия (хотя Кохави и др. утверждают, что этот эффект нельзя назвать широко распространенным). Однако преимущества этого типа тестирования достигаются за счет увеличения сложности организации, проведения и анализа тестирования. Его проведение рационально только при условии достаточно высокого трафика для сохранения статистической мощности.

БАЙЕСОВСКИЕ БАНДИТЫ

A/B-тестирование, описанное в этой главе, более широко распространено и популярно на практике. Оно осуществляется в рамках классического, или частотного, статистического подхода. Однако существует еще один подход, который набирает популярность в последние годы благодаря стремительному развитию вычислительных технологий, — это байесовская статистика[1].

В рамках частотного подхода стартовая точка — формулировка гипотезы, например «CTR в контрольной группе равен CTR в тестовой группе». Вы собираете данные и задаете вопрос: «Какова вероятность получения тех же самых (или более значимых) результатов при многократном повторении эксперимента, если эта гипотеза верна?» При этом по умолчанию предполагается, что внешние условия не меняются, то есть мы в вероятностном смысле делаем выводы из распределения, но само распределение и его параметры со временем остаются неизменными.

В рамках байесовского подхода все по-другому. Стартовой точкой служит предпосылочное убеждение. Что мне известно об этой системе? Возможно, ранее вам еще не приходилось тестировать подобные характеристики, и тогда вы начинаете с простой догадки. Возможно, наоборот, у вас уже был опыт, и вы можете использовать полученную ранее информацию как основу. Хотя фактически предпосылочные убеждения играют не настолько важную роль, так как со временем вы будете обновлять и изменять их по мере получения новых доказательств. Даже если изначально они были ошибочными, постепенно они будут меняться и в большей мере отражать действительность. Это ключевое отличие от частотного подхода: любая новая информация — просмотр, продажа или переход по ссылке — становится дополнительным

[1] URL: http://www.austincc.edu/mparker/stat/nov04/talk_nov04.pdf.

доказательством, которое следует включать в базу знаний. Это итеративный подход. Более того, в его рамках не стоит вопрос «Есть ли различие между сравниваемыми вариантами?», вместо этого задают другой вопрос: «Что эффективнее: контрольный параметр или тестовый?» И это то, что хочет знать бизнес.

Если вас заинтересовал термин «бандит», то он появился по аналогии с игровыми автоматами, которые иногда еще называют «однорукими бандитами». Суть в том, что мы имеем дело со множеством «бандитов» (один контрольный и множество тестовых), у каждого из которых разная частота выигрыша (внутренний коэффициент CTR). Нам нужно выявить лучшего «бандита» (самый высокий коэффициент CTR), но сделать это мы можем только с помощью серии нажатия рычага (показов). Каждый бандит выдает выигрыш случайным образом, а значит, нам нужно сбалансировать нажатие рычагов у потенциально менее перспективных «бандитов», чтобы получить дополнительную информацию, по сравнению с нажатием рычага только у того автомата, который мы считаем самым перспективным, чтобы максимизировать получение выигрыша.

Со временем система будет менять соотношение пользователей, которые получают более эффективную характеристику. Грубо говоря, тестирование может начаться с соотношения 50/50. Предположим, что тестируемая характеристика действительно *очень* эффективна (мы наблюдаем гораздо больше переходов), тогда система снижает пропорцию посетителей, которые пользуются контрольной характеристикой, и увеличивает пропорцию тех, кто пользуется тестируемой характеристикой. Теперь соотношение составляет 40% (контрольная группа) и 60% (тестовая). Мы продолжаем наблюдать значительный положительный эффект, и процентное соотношение вновь корректируется: 30% (контрольная группа) и 70% (тестовая) и так далее. У этого подхода два очевидных преимущества. Во-первых, нет необходимости проводить анализ, чтобы понять, какой вариант лучше, — можно просто оценить относительную пропорцию. Во-вторых, поскольку более эффективная характеристика применяется дольше, у нас есть возможность сразу же воспользоваться этим преимуществом. (В терминах статистики, нам не придется сожалеть об упущенной выгоде за период проведения эксперимента, когда у нас все еще действовала менее эффективная характеристика.)

В отличие от частотного подхода, здесь имеется возможность добраться до максимальных значений и наблюдать за изменением системы

на протяжении времени. Здесь нет фиксированного периода проведения эксперимента: он может длиться бесконечно. Фактически мы можем добавлять характеристики, исключать их, изменять. В рамках частотного подхода это было бы невозможно. Можно продолжать эксперимент или установить ограничивающий критерий: например, если эффективность тестируемой характеристики превышает 5% по сравнению с контрольной характеристикой, 100% трафика переключается на нее.

Разумеется, я опустил множество математических деталей, самая главная из которых — правило обновления, или то, как происходит изменение степени вероятности. Фактически система разработана таким образом, что проходит этап *изучения,* на котором вы пробуете все разные контрольные и тестовые характеристики с относительной частотностью, а затем этап *использования*, на котором вы активно используете наиболее эффективную на данный момент характеристику (и минимизируете сожаление). При байесовском подходе наблюдаются те же самые проблемы, что и при частотном подходе: положительный результат тестируемой характеристики может быть как ее эффектом, так и делом случая. Если результат был случайным, то дальнейшее использование этой характеристики, скорее всего, приведет к снижению коэффициента CTR, и пропорция тестовой группы будет скорректирована в сторону снижения по правилу обновления. Это означает, что такая система не в состоянии гарантировать системное повторение одного и того же опыта для каждого пользователя или хотя бы для пользователей, посещающих сайт повторно.

Байесовский подход набирает популярность, хотя и медленно. Гораздо сложнее объяснить неспециалистам принцип работы системы, но зато интерпретировать результаты проще. В отличие от частотного подхода, нет необходимости устанавливать продолжительность тестирования — вместо этого можно определить ограничивающий критерий, что с точки зрения бизнеса сделать легче. Мне интересно, можно ли считать одной из причин медленного внедрения этого подхода сам алгоритм, который производит модификации со временем и определяет, какую версию сайта увидит пользователь, — ведь фактически всем управляет байесовское правило обновления. В компании должна быть очень хорошо развита культура работы с данными, чтобы сотрудники могли доверять этому процессу. К сожалению, для многих компаний эта система не более чем волшебный черный ящик.

Влияние на корпоративную культуру

Мы рассмотрели технические аспекты проведения тестирования для достижения максимального эффекта, и теперь я хочу остановиться на вопросах влияния этого процесса на корпоративную культуру компании.

Скотт Кук, основатель компании Intuit, считает, что A/B тестирование сдвигает фокус с «принятия решений на основе убеждения» на «принятие решений на основе экспериментов»[1]. Эта философия не подпитывает ничье эго. Теперь правила игры задают не HiPPO (highest paid person's opinion, то есть «мнение самого высокооплачиваемого сотрудника»): происходит демократический сдвиг от принятия решений на высшем уровне к генерированию гипотез на уровне операционном. Скотт Кук полагает, что таким образом компания поощряет даже сотрудников на незначительных должностях тестировать свои лучшие идеи. У сотрудников появляется больше эффективных идей, чувство сопричастности, собственности и вовлеченности. Как я призываю в одном из постов в блоге (который фактически лег в основу этой книги), «дайте слово молодым специалистам»[2].

Сирокер и Кумен утверждают, что подобный подход позволяет компании раздвинуть границы и стать более инновационной. «Он убирает требование, по которому все вовлеченные в процесс должны знать всё. Когда сотрудники могут спокойно сказать: "Я не знаю, но давайте проведем эксперимент", — они больше склонны принимать на себя ответственность и рисковать делать вещи, выходящие за рамки нормы». Скотт Кук полностью с этим согласен. По его словам, когда люди экспериментируют, «они чаще удивляются, а удивление — источник инноваций. Человек удивляется, только когда делает что-то и получает результат, отличающийся от его ожиданий. Так что чем скорее вы начнете экспериментировать, тем скорее начнете удивляться и открывать для себя то, чего не знали раньше».

Кроме того, Сирокер и Кумен полагают, что время рабочих встреч можно сократить. Они цитируют Джарреда Колли, бывшего старшего менеджера по маркетингу продукта в компании Rocket Lawyer: «Если раньше сотрудники могли часами с пеной у рта спорить, какой заголовок или какое изображение лучше использовать, сейчас необходимость

---

[1] URL: https://www.fastcompany.com/3020699/bottom-line/why-intuit-founder-scott-cook-wants-you-to-stop-listening-to-your-boss.

[2] URL: http://www.p-value.info/2013/04/how-do-you-create-data-driven.html.

в этих обсуждениях отпала: мы просто всё тестируем и точно знаем, что лучше». Опять-таки, больше не происходит столкновения самомнений, больше не надо изобретать теории, вместо того чтобы сосредоточиться на идеях, которые могут просто работать и приносить пользу. Большинство идей не оказывают никакого влияния или способны сделать только хуже, но, чтобы добиться значительного эффекта, достаточно всего одного или двух удачных попаданий. Вспомните о дополнительных 57 млн долл., которые стали результатом оптимизации подписной страницы кампании Барака Обамы. Это огромная рентабельность от затраченных усилий. Но даже эта цифра меркнет в сравнении с той пожизненной ценностью, которую принесла компании Amazon.com идея Грега Линдена с рекомендательным сервисом при оформлении заказа. Недавно разработчики поисковой системы Bing тестировали, улучшатся ли результаты, если увеличить количество ссылок в рекламных объявлениях. В результате теста выяснили, что две и более ссылок лучше, чем одна, и предположительно это принесло сервису 100 млн долл. *ежегодно*[1]. Это не случайная удача: одновременно проводили 300 тестов в день. Google постоянно проводит тысячи экспериментов. Чтобы достигать результатов, нельзя останавливаться. Есть даже шутка, что «A/B-тестирование» на самом деле расшифровывается как «тестирование абсолютно всегда».

[1] URL: https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/parmyolson/2015/01/21/jawbone-guinea-pig-economy/&refURL=&referrer=.

ГЛАВА 9

Принятие решений

В значительном количестве компаний топ-менеджмент принимает решения за закрытыми дверями, не привлекая к этому процессу особого внимания, чтобы избежать ответственности, если эти решения окажутся неверными. Такое положение дел вызывает тревогу.

Аналитическое подразделение группы компаний Economist Group, издателя журнала *Economist*[1]

Нет никакой мистики в процессе принятия решений. Обучиться этому навыку может каждый.

Сидни Финкельштайн (там же)

Осторожно: сомнительная шутка.

Какое животное лучше всех управляется с данными? Гадюка[2] (можете разочарованно выдохнуть). А какое животное данные волнуют меньше всего? Гиппопотама (HiPPO). И здесь все гораздо серьезнее. HiPPO — аббревиатура от highest paid person's opinion, то есть «мнение самого высокооплачиваемого сотрудника» (рис. 9.1). Этот термин ввел в употребление Авинаш Кошик для обозначения концепции, полностью противоположной управлению на основе данных. Каждый из нас сталкивался с такими людьми — это эксперты с многолетним опытом. Им наплевать на данные, особенно когда те идут вразрез с их персональным мнением, и они всегда придерживаются своего плана, потому что знают лучше. Кроме того, «они здесь начальники», как объясняет Financial Times[3]:

---

[1] Economist Intelligence Unit, Decisive Action: how businesses make decisions and how they could do it better (London: Economist Intelligence Unit, 2014). URL: http://www.datascienceassn.org/sites/default/files/Decisive%20Action%20-%20How%20Businesses%20Make%20Decisions%20and%20How%20They%20Could%20do%20it%20Better.pdf.

[2] Игра слов строится на многозначности английского слова adder, которое используется в оригинале и имеет следующие значения: 1) гадюка, змея; 2) счетный прибор. *Прим. перев.* Подробнее см. по ссылке: https://en.wikipedia.org/wiki/Adder.

[3] Lynch M. Is your HiPPO holding you back? Financial Times, September 14, 2009. URL: https://www.ft.com/content/62f37a4a-931c-11de-b146-00144feabdc0.

> Так называемые HiPPO могут быть крайне опасны для бизнеса, поскольку принимают решения в лучшем случае на основе неверной интерпретации данных, а в худшем — на основе беспочвенных догадок. Они не прибегают к инструментам бизнес-аналитики, чтобы понять поведение клиентов и оценить причины («как», «когда», «где» и «почему»), которые обусловливают это поведение. Подход HiPPO может стать губительным для компании.

Эта глава посвящена тому звену в аналитической цепочке ценностей, которое, вероятно, обычно обсуждается меньше всего, — непосредственно процессу принятия решений. В компании может осуществляться качественный и своевременный сбор необходимых данных, может быть опытный специалист по работе с этими данными, который составляет полезные отчеты и модели и формулирует важные выводы и рекомендации. Но если эти отчеты пылятся на полках или руководитель принимает решения по наитию, независимо от того, что показывают данные, то это все лишено смысла.

Рис. 9.1. Решения должны приниматься на основе данных, а не мнения HiPPO

Иллюстрация Тома Фишбурна. Воспроизводится с разрешения

В этой главе мы рассмотрим ряд вопросов, которые касаются процесса принятия решений. Во-первых, остановимся на том, как осуществляется этот процесс. Обычно решения принимаются на основе данных или

на основе мнения HiPPO? Я объясню, что на самом деле подразумевает термин «управление на основе данных» и как он соотносится с другими похожими терминами: «информация на основе данных» и «влияние на основе данных». Далее мы подробно изучим, что может затруднять процесс принятия решений, и коснемся таких аспектов, как данные, корпоративная культура и когнитивные искажения (иррациональное или нелогичное мышление). Обозначив спектр проблем и рискуя вогнать читателей в депрессию, я переключусь на способы решения этих проблем и рекомендации по повышению качества процесса принятия решений на основе фактов. Все это я буду делать в рамках поведенческой модели Фогга[1].

Как принимают решения?

Здесь не все так просто. Многие компании искренне верят, что у них процесс принятия решений происходит на основе данных, но, к сожалению, интуиция по-прежнему правит бал. Вот некоторые факты: интуиция и персональный опыт заняли первые две строчки в рейтинге факторов, на основе которых топ-менеджмент принимает решения, согласно отчету компании Accenture в 2009 году (n = 600; рис. 9.2).

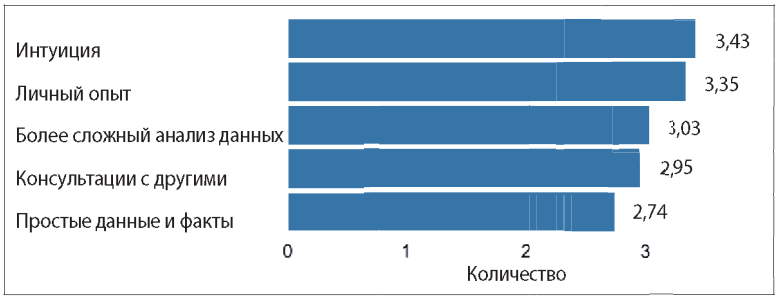


Рис. 9.2. Факторы, на основе которых топ-менеджмент принимает решения

Подготовлено по рис. 5 отчета Analytics in Action: Breakthroughs and Barriers on the Journey to ROI компании Accenture

Источник: https://www.accenture.com/us-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Technology_6/Accenture-Analytics-In-Action-Survey.pdf

В исследовании 2014 года, которое проводило аналитическое подразделение журнала Economist, на основе опроса 1135 руководителей высшего звена получилась аналогичная картина (рис. 9.3): интуиция

[1] Поведенческая модель Фогга (Fogg Behavior Model), или FBM, — модель, согласно которой поступок — это следствие трех факторов: мотивации, способностей и стимула.

(30%) и опыт (28%) в совокупности оставили далеко позади аналитический подход (29%)[1].

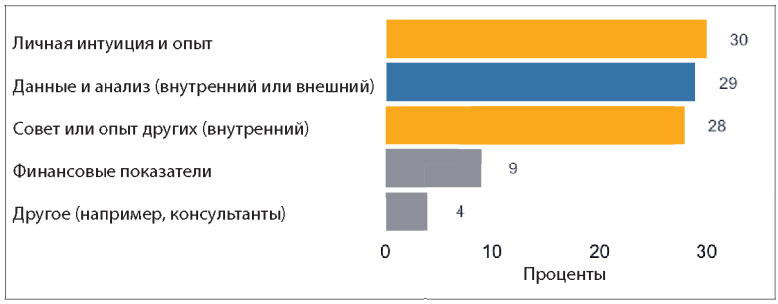


Рис. 9.3. На какой из следующих факторов вы опирались в большей мере при принятии последнего серьезного бизнес-решения?

По результатам другого опроса, в котором приняли участие более 700 топ-менеджеров, 61% респондентов заявили, что при принятии решений следует прислушиваться к практическому опыту, а не к цифрам, а 62% опрошенных уверены, что часто необходимо и даже предпочтительно полагаться на интуицию и «мягкие» факторы[2].

Наконец, в опросе IBM с участием 225 руководителей по всему миру интуиция и опыт вновь возглавляют список[3]. См. табл. 9.1.

Таблица 9.1. В какой степени вы руководствуетесь следующими факторами при принятии бизнес-решений?

Фактор	Часто	Всегда	Итого
Личный опыт и интуиция	54%	25%	79%
Аналитические данные	43%	19%	62%
Коллективный опыт	43%	9%	52%

Как следует из результатов четырех исследований, картина примерно одинаковая.

Тем не менее мне удалось найти один отчет, где подход на основе данных обошел другие (рис. 9.4). Это еще один опрос аналитического подразделения журнала Economist от 2014 года ($n = 174$)[4].

[1] URL: http://www.pwc.com/us/en/advisory-services/data-possibilities/big-decision-survey.html.

[2] URL: https://www.gyro.com/onlyhuman/gyro-only-human.pdf.

[3] URL: http://www-05.ibm.com/de/services/bao/pdf/gbe03211-usen-00.pdf.

[4] Отчет размещен на сайте http://www.eiu.com/.

Рис. 9.4. Какой из следующих пунктов лучше всего описывает ваш личный подход при принятии важных управленческих решений?

См. также рис. 7 отчета Analytics: a blueprint for value, IBM

Источник: http://www-935.ibm.com/services/us/gbs/thoughtleadership/ninelevers/

Как можно объяснить подобные результаты? Почему субъективному опыту и интуиции отдают предпочтение по сравнению с объективным аналитическим подходом? Если не вдаваться в детали, можно выделить три фактора: сами данные, корпоративная культура компании и, наконец, человеческий мозг. Чуть позже я подробнее остановлюсь на каждом из них, чтобы показать некоторые глубинные причины, обусловливающие процесс принятия решений. После этого мы обсудим некоторые возможные решения и подходы.

Прежде всего давайте посмотрим, можем ли мы в принципе быть объективными при принятии решений. Разве мы не всегда прислушиваемся к своим внутренним ощущениям? Что фактически мы имеем в виду на данном этапе аналитической цепочки, говоря об управлении на основе данных?

УПРАВЛЕНИЕ, ИНФОРМИРОВАНИЕ ИЛИ ВЛИЯНИЕ НА ОСНОВЕ ДАННЫХ?

На протяжении всей книги я употребляю термин «управление на основе данных». В главе 1 я представил общий обзор этой концепции и использую ее впоследствии относительно данных. Тем не менее имеет смысл подробнее остановиться на понятии «управление». Насколько мы действительно управляем на основе данных? Может быть, другие понятия, такие как «получение информации на основе данных» или «стимулирование влияния на основе данных», более уместны?

Скотт Беркун затрагивает некоторые действительно важные аспекты в своем посте, озаглавленном The Dangers of Faith in Data («Опасность веры в данные»)[1]. Он утверждает: «Данные не могут управлять. Они не наделены сознанием — это просто набор мертвых цифр. У данных нет интеллекта, следовательно, они неспособны ничем управлять». Думаю, этот пост может послужить хорошей темой для обсуждения с коллегами из аналитического отдела. Сам пост, очевидно, вызовет жаркие споры, но в нем есть некоторые весьма ценные идеи, достойные того, чтобы над ними поразмышляли.

Если управление у вас ассоциируется с управлением автомобилем — данные говорят повернуть налево, и вы поворачиваете налево, — то в большинстве случаев это не сработает практически ни в какой компании, если только она не руководствуется аналитикой очень высокого уровня (главы 2 и 5). Если в своей работе вы сталкиваетесь с одними и теми же регулярно повторяющимися ситуациями и у вас разработаны действительно качественные прогностические модели, тогда у вас непременно должны быть решения на основе данных, которые принимаются автоматически. Например, рассмотрим ситуацию пополнения товарных запасов в производственном процессе. Эту функцию можно автоматизировать: вы разрабатываете прогностический алгоритм, который отслеживает уровень продаж и запасы на складе и отправляет заказы на пополнение запасов так, чтобы не образовывалось дефицита товара, но чтобы уровень запасов был минимальным. Или возьмем, например, автоматизированные торговые системы, в которых алгоритмы независимо продают настоящий товар за настоящие деньги. В подобных сценариях человек, принимающий решения, фактически оказывается над системой, а решения, влияющие на компанию, принимаются автоматически на основе данных и алгоритмов. Я согласен со Скоттом, что в большинстве случаев понятие «управление на основе данных» подразумевает несколько иное.

Тем не менее это не единственное значение термина «управление». Одно из его зафиксированных словарных значений — «причина (нечто абстрактное), обусловливающая что-то происходящее». Пример употребления этого значения: «На протяжении ряда лет потребитель управляет развитием экономики». Очевидно, что потребители не сидят в своих гостиных с пультами в руках и не контролируют такие показатели, как, скажем, инфляция, при этом их поведение действительно фактор развития экономики. Уровень потребления

[1] URL: http://scottberkun.com/2013/danger-of-faith-in-data.

населения, объем кредитных обязательств и сбережений — все эти факторы, в совокупности с интерпретацией этих данных главой Федеральной резервной системы, формируют экономику. Например, значения таких ключевых показателей, как уровень безработицы, потребительские расходы и владение недвижимостью, мотивировали Бена Бернанке[1] сохранить процентные ставки на низком уровне для стимулирования экономического роста. Его никто не заставлял это делать, данные не приставляли пистолет к его виску, но направленность этих основных показателей плюс практический опыт и знания в области кредитно-денежной политики действительно обусловили его решения. (Аналогичным образом, я уверен, что в исследованиях, о которых говорилось чуть выше, данные не противопоставлялись интуиции — скорее, речь шла об интуиции в отсутствие любых актуальных данных. Именно это противопоставлялось аналитическому подходу, при котором осуществлялся сбор и анализ данных в сочетании с опытом и знаниями руководителя.) Я склонен понимать под управлением на основе данных именно такое сочетание. Скотт продолжает: «В лучшем случае можно стремиться к тому, чтобы данные оказывали влияние на принятие решений, то есть чтобы опытные руководители располагали адекватными данными, на которые они могут опираться в поисках ответов на правильные вопросы о том, что и насколько эффективно они делают и что, возможно, им следует делать в будущем». Я полностью согласен с этой точкой зрения. По моему мнению, термин «управление на основе данных» можно использовать именно в этом смысле.

Кнапп и др. предпочитают термин «информация на основе данных», по крайней мере, в контексте образовательного управления:

> Мы считаем концепцию управления с информацией на основе данных более полезной… В этом случае горизонт мышления и действий расширяется в двух направлениях. Во-первых, появляется возможность избежать ощущения, что данные «управляют» действиями (это отсыл к примеру с управлением автомобилем). Во-вторых, эта концепция предполагает, что данные более полезны для практики управления, чем для принятия решений как таковых…

[1] Бен Шалом Бернанке (Ben Shalom Bernanke; р. 1953) — американский экономист, председатель Совета экономических консультантов при Белом доме. Председатель совета управляющих Федеральной резервной системы США с февраля 2006 до февраля 2014 года. *Прим. ред.*

> Данные в большей степени задают вопросы и стимулируют размышления, чем указывают на конкретные варианты решения проблемы[1].

Иными словами, авторы выступают за то, что данные обеспечивают информацию для принятия решений (в том смысле, в котором Скотт говорил о *влиянии* на основе данных), а также помогают ставить вопросы и информируют о том, что происходит в компании, например каковы ключевые показатели эффективности, отчеты и оповещения. Они также цитируют Бернхардта: «Настоящее принятие решений на основе данных лишь частично зависит от данных. В процессе принятия решений основная роль принадлежит четкому ви́дению, которое разделяют все, и управлению».

Все три термина имеют смысл и право на существование. «Влияние», на мой взгляд, — самый слабый и пассивный из них, а «управление» — самый сильный и активный. Независимо от того, какой из этих терминов объективно лучше, арбитром в этом споре стала Google. На момент написания книги по ключевому слову data-influenced («влияние на основе данных») поисковая система выдавала 16 тыс. результатов, по ключевому слову data-informed («информирование на основе данных») — 170 тыс. результатов, и по ключевому слову data-driven («управление на основе данных») — 11,5 *млн* результатов. Таким образом, правильно это или нет, но именно термин «управление на основе данных» завоевал наибольшую популярность, получил наиболее широкое распространение и используется в этой книге.

Что осложняет процесс принятия решения?

В этой части мы изучим факторы, осложняющие процесс принятия решения и стимулирующие принятие решения на основе внутренних ощущений.

ДАННЫЕ

Как уже отмечалось ранее (см. главу 2), данные должны отличаться своевременностью, адекватностью и достоверностью. В противном случае у человека, принимающего решение, весьма ограничены варианты действий. Он может отложить принятие решения, постараться

[1] URL: https://www.naesp.org/resources/2/Research_Roundup/2008/RR2008v24n3a3.pdf.

собрать больше данных или принять решение на основе имеющихся в его распоряжении данных и инструментов, что обычно сводится к одному только практическому опыту.

Какие проблемы могут возникнуть с данными?

Качество данных и недостаток доверия к ним

Возвращаясь к результатам одного из исследований, о которых говорилось ранее в этой главе[1], при принятии решений сами данные могут представлять настоящую проблему: «Главное препятствие для более эффективного использования этого актива при принятии решений — качество данных, их точность и полнота».

По данным Harvard Business Review, «51% респондентов располагали необходимой информацией, чтобы чувствовать себя уверенно при принятии деловых решений за последние шесть месяцев. Эта группа имеет неоспоримые преимущества: они чувствуют себя увереннее при необходимости принять решение с высокой степенью риска и ощущают себя готовыми своевременно принимать серьезные бизнес-решения»[2]. Это отлично, но как насчет тех 49% респондентов, которые не располагают необходимыми им данными, чтобы чувствовать себя уверенно? По результатам другого исследования, каждый третий руководитель принимал важные решения, обладая неполной информацией или информацией, которой он не доверял[3]. Исправить сложившуюся ситуацию можно только с внедрением принципов лидерства на основе данных, которые подразумевают инвестиции в развитие управления на основе данных и программы повышения качества данных.

Объем

Для других проблема заключается не в недостатке данных, а, наоборот, в их избытке. Они не могут справиться с чрезмерным объемом. В том же самом исследовании HBR говорится: «Более половины респондентов указывали, что объем как внутренних, так и внешних данных, необходимых для принятия решения, увеличивается быстрее, чем компания способна обработать». В этом случае сосредоточьтесь на выборках,

---

[1] URL: http://www.pwc.com/us/en/advisory-services/data-possibilities/big-decision-survey.html.

[2] URL: https://hbr.org/resources/pdfs/tools/HBR_Qlik_Report_May2014.pdf.

[3] URL: http://www-05.ibm.com/de/services/bao/pdf/gbe03211-usen-00.pdf.

сокращайте объем до самого важного, агрегируйте и автоматизируйте, при необходимости наймите дополнительных специалистов по сбору и обработке данных.

Разделение сигнала и шума

Большой объем оборачивается и другими проблемами. Чем больше у вас данных, тем больше сигналов, но и информационного шума тоже больше. Выделить то, что действительно важно, становится сложнее. Особенно это касается больших данных, где фиксируется и сохраняется абсолютно все. Объем нужных данных размывается, и аналитикам бывает сложно отделить сигнал от информационного шума.

В этом случае вместо общего копания данных в надежде наткнуться на что-то значимое и важное может помочь четкая постановка вопроса. Однако даже тогда бывает непросто отделить зерна от плевел. «Слишком много доказательств может быть так же плохо, как и слишком мало», — говорит Джерард Ходкинсон, профессор стратегического менеджмента Школы бизнеса Университета Уорвика (отчет Decisive Action).

Пол Андриссен (1988)[1] провел эксперимент со студентами, изучающими бизнес в Массачусетском технологическом институте. Участников эксперимента разделили на две группы и предложили сформировать собственный портфель капиталовложений. У первой группы был ограничен доступ к информации, им было известно только о колебаниях цен на активы[2]. А у второй группы доступ к информации был неограниченным: они не только следили за изменением цен на акции, но и могли получать другие финансовые новости из газет, ТВ, радио и так далее. Обеим группам предложили принять участие в биржевых торгах. У кого результаты были лучше? Возможно, вы удивитесь, но результаты первой группы (с ограниченным доступом к информации) оказались лучше в два раза. Участники второй группы получали гораздо больше сигналов, слухов, сплетен и уделяли слишком много внимания тому, что того не стоило. Они искали сигналы в информационном шуме и заключали больше сделок. (Этот эффект носит название «склонность к поиску информации» — information bias.) Например, трейдеры фиксируются на недавних максимальных или минимальных

[1] Mussweiler T. and Schneller K. “What goes up must come down” — how charts influence decisions to buy and sell stocks, Journal of Behavioral Finance 4, no. 3 (2003): 121–130.

[2] URL: https://www.fastcompany.com/45655/too-much-information.

значениях стоимости акций, которые по определению являются экстремумами, и используют их как якоря (подробнее об этом далее). Соответственно, это стимулирует их продавать или покупать активы.

Если вас интересуют примеры из других областей, помимо финансовой, я рекомендую вам книгу Барри Шварца Paradox of Choice[1] (Harper Perennial). В ней описывается достаточно случаев, когда избыток вариантов выбора и информации способен вызвать «аналитический паралич».

Это лишь некоторые проблемы, которые могут возникнуть с данными. По результатам большого опроса руководителей, «менее 44% сотрудников знают, где найти информацию, необходимую им в повседневной работе». Но даже если им известен источник, где искать, данных может быть недостаточно или они невысокого качества. Неудивительно, что, «если руководители стоят перед выбором воспользоваться достаточно хорошими данными сейчас или получить более качественные данные, но позже, большинство из них остановятся на первом варианте, так как уверены, что смогут восполнить пробелы благодаря своему опыту и знаниям»[2]. В этом и заключается проблема.

КОРПОРАТИВНАЯ КУЛЬТУРА

Еще один аспект, влияющий на процесс принятия решений, — сложившаяся в компании корпоративная культура. (Корпоративная культура — вероятно, наиболее значимый фактор в компании с управлением на основе данных. Подробнее мы поговорим о ней в главе 10.)

Ценность интуиции

Руководители высшего звена, как правило, отличаются от рядовых сотрудников способностью мыслить стратегически. Часто под этим подразумевается их способность создать ви́дение, воплотить его в жизнь, добиться поставленной цели, справляясь со всеми препятствиями на пути, независимо от того, что говорят данные. Топ-менеджер должен обладать хорошей интуицией. Часто он получает место именно за свою интуицию. Да что там говорить, биография Джека Уэлча, легендарного бывшего генерального директора компании General Electric,

---

[1] Издана на русском языке: Шварц Б. Парадокс выбора. Почему «больше» значит «меньше» М. : Добрая книга, 2005.

[2] Shah S., Horne A. and Capella J. Good data won't guarantee good decisions, Harvard Business Review 90, no. 4 (2012): 23–25.

называется Straight from the Gut[1] (в дословном переводе «На основе шестого чувства»). (Но нужно отдать ему должное, Уэлч умеет работать с данными и продвигал концепцию «Шесть сигм».)

Неумение работать с данными

Серьезная проблема заключается в том, что многие топ-менеджеры не умеют работать с данными. То есть прошел уже не один год, а может быть, даже не одно десятилетие с тех пор, как они изучали такую дисциплину, как статистика (если изучали в принципе). Эта дисциплина не входит в программу MBA, и коучи ей тоже не обучают. Эта статистическая безграмотность весьма некстати, так как именно руководители становятся последней линией обороны. Именно руководитель получает набор агрегированных данных, интерпретирует выводы и рекомендации аналитиков, оценивает убедительность доказательств, степень риска и влияние тех шагов, которые должны продвинуть компанию вперед.

Два названных фактора в совокупности свидетельствуют о том, что HiPPO — это не такое уже редкое явление, и часто эти люди обладают определенной властью в компании.

Отсутствие прозрачности

Если объединить три этих фактора: приоритет интуиции, неумение работать с данными и неподотчетность, — получится смертельная комбинация. В ходе одного из опросов (рис. 9.5) 41% респондентов сказали, что люди, не умеющие принимать решения, не смогут продвигаться в их компании по карьерной лестнице, и это означает, что в большинстве случаев (59%) такие люди *растут* по карьерной лестнице. Кроме того, 19% респондентов указали, что в их компании люди, принимающие решения, не отчитываются за эти решения. А 64% опрошенных заявили, что информация о том, кто принимал конкретное решение, известна только топ-менеджменту.

Это означает, что качество решений половины руководителей никак не оценивается. Кроме того, они не отчитываются за принятые решения. Если у такого руководителя нет навыка работы с данными, что удерживает его от того, чтобы превратиться в HiPPO? Подотчетность должна быть и на уровне аналитической работы с данными. (Вспомните

[1] Издана на русском языке: Уэлч Дж., Бирн Дж. Джек Уэлч. История менеджера. М. : Манн, Иванов и Фербер, 2012.

слова Кена Рудина: «Смысл аналитики в оказании влияния… В нашей компании [Zynga], если вы провели блестящее исследование и сделали потрясающие выводы, но ничего не изменилось, результативность вашей работы равна нулю»). Аналитики должны убеждать руководство в своих выводах и приводить веские доказательства. Они должны предоставлять достоверную информацию о размере выборки, относительной величине погрешности, доверительных интервалах. Более того, обо всем этом они должны говорить языком, понятным руководителю.

Рис. 9.5. Как в вашей компании люди, принимающие решения, отвечают за них?

Источник: Decisive Action: how businesses make decisions and how they could do it better, аналитическое подразделение журнала Economist.
URL: http://thedecisionengineer.com/decisive-action-business-growth/

КОГНИТИВНЫЕ БАРЬЕРЫ

Мы обсудили такие факторы, влияющие на принятие решений, как недостаток прозрачности, нехватка навыков и приоритет интуиции в рамках корпоративной культуры компании. Есть еще один огромный барьер, препятствующий эффективному принятию решений и поддерживающий (плохую) интуицию, — наш мозг.

Горькая правда в том, что мы принимаем решения, далекие от идеальных. Мы не всегда решаем проблемы наиболее объективным образом, часто держимся за устаревший опыт и зацикливаемся на ненужных деталях, что ведет к нерациональному мышлению. Эти влияния и механизмы носят название когнитивных искажений. Для знакомства с темой рекомендую книгу Рольфа Добелли The Art of Thinking Clearly или список в «Википедии»[1].

[1] URL: https://en.wikipedia.org/wiki/Category:Cognitive_biases.

В человеческом сознании процесс принятия решений происходит двумя основными способами: быстро, непреднамеренно, неосознанно (лауреат Нобелевской премии Даниэль Канеман назвал это системой 1) и медленно и намеренно (система 2). Система 1 — это наше «шестое чувство», интуиция, в то время как система 2 — это наше сознание, мы пользуемся ею для тщательного обдумывания и глубокого математического анализа.

Давайте посмотрим, почему мы не можем всегда доверять интуиции[1].

Мы не отличаемся постоянством

Одни и те же доказательства в разное время приводят нас к отличающимся друг от друга заключениям. Более того, если разные люди получают одни и те же доказательства, они делают разные выводы[2].

Мы помним то, что не происходило

Интуиция человека основана на подсознательном сборе информации, но при этом не все полученные данные достоверны. В увлекательной статье об очевидцах, вспоминающих то, чего никогда не происходило, которая была опубликована в New York Times[3], авторы предполагают, что «память человека хранит обрывки правды, окруженные дырами, которые человек заполняет собственными догадками и убеждениями».

Мы не настолько компетентны, как нам кажется

По Канеману, человеку свойственна «иллюзия правильности». Вот простой пример. Попробуйте ответить как можно быстрее.

---

[1] Изложенное дальше преимущественно основывается на книге Даниэля Канемана Thinking, Fast and Slow (Farrar, Straus and Giroux, 2011) (издана на русском языке: Каneман Д. Думай медленно… решай быстро. М. : АСТ, 2016). Настоятельно рекомендую эту книгу к прочтению. Если у вас нет времени прочитать книгу полностью, прочитайте хотя бы отличную обзорную статью: Kahneman D. and. Klein G. Conditions for intuitive expertise: A failure to disagree, American Psychologist 64, no. 6 (2009): 515–526. А также McAfee A. The Future of Decision Making: Less Intuition, More Evidence, Harvard Business Review, January 7, 2010. URL: https://hbr.org/2010/01/the-future-of-decision-making.

[2] Frick W. What to Do When People Draw Different Conclusions From the Same Data. Harvard Business Review, March 31, 2015. URL: https://hbr.org/2015/03/what-to-do-when-people-draw-different-conclusions-from-the-same-data.

[3] URL: https://www.nytimes.com/2015/05/15/nyregion/witness-accounts-in-midtown-hammer-attack-show-the-power-of-false-memory.html?smprod=nytcore-iphone&smid=nytcore-iphone-share&_r=0.

Бейсбольная бита и мяч вместе стоят 1,1 долл.
Бита стоит на 1 долл. дороже мяча.
Сколько стоит мяч?

Большинство людей, включая меня, отвечают 0,1 долл. — и ошибаются. Правильный ответ — 0,05 долл. Ответ нашей интуитивной системы 1 неверный, а система 2 слишком ленива, чтобы это проверить. Тем не менее, если сразу включить рациональное мышление системы 2, можно легко найти правильный ответ: цена биты — 1,05 долл., а цена мяча — 0,05 долл., а также проверить его правильность: 1,05 долл. + 0,05 долл. = 1,1 долл. и 1,05 долл. – 0,05 долл. = 1 долл. (Если вы тоже дали неверный ответ, не расстраивайтесь: уровень ошибки среди студентов престижнейших университетов США, таких как МТИ, Принстон и Гарвард, составил 50%, а в менее престижных университетах приблизился к 90%.)

Мы с трудом отказываемся от устаревшей информации

Человек усваивает факты, строит на их основе ментальные модели, а когда получает данные, противоречащие первоначальным фактам, с трудом воспринимает новую информацию и неохотно меняет свою модель. Брендан Найхен и Джейсон Райфлер из Дартмурского колледжа провели ряд исследований, в которых участникам предлагали прочитать фальшивую газетную статью, содержавшую либо ложное заявление политика, либо ложное заявление и его опровержение. Они обнаружили, что «те участники, которые получили нежелательную информацию [то есть с опровержением, которое шло вразрез со сложившимся у них убеждением], не смогли сразу отказаться от своей точки зрения. Вместо этого они начинали отстаивать ее более активно, это проявление так называемого «эффекта обратного результата»[1]. Авторы исследования процитировали Марка Твена: «Неприятности доставляет не то, чего вы не знаете, а то, что вы знаете наверняка и что оказывается неверным». Иными словами, дезинформация более опасна, чем односторонний взгляд на вещи: дезинформация прилипчива. Как сказал на конференции 2014 Strata+Hadoop World в Нью-Йорке автор книги The Hidden Brain (Spiegel & Grau) Шанкар Вендантам, «фактически знания никак не влияют на нашу дезинформацию, а в некоторых случаях только усугубляют ее»[2].

[1] URL: http://www.dartmouth.edu/~nyhan/nyhan-reifler.pdf.

[2] URL: https://www.youtube.com/watch?v=7mpe6luA5Os.

Мы фиксируемся на не имеющих значения данных

Если вам доводилось покупать автомобиль, то, скорее всего, сначала вы узнали его официальную цену, а затем, если вы человек рациональный, вероятно, начали торговаться с менеджером, который долго ломался, мямлил, ходил «поговорить с боссом», но наконец согласился дать вам скидку. Добившись снижения цены, вы, возможно, порадовались, что заключили выгодную сделку. Но правда в том, что «официальная» цена — это полная ерунда. Это психологическая уловка, чтобы заставить вас мыслить относительными категориями и сравнивать полученное предложение с более высоким, вместо того чтобы сосредоточиться на абсолютном объеме или другом прямом доказательстве. Ваше внимание пытаются зафиксировать на этом значении, которое воспринимается как ориентир.

В данном случае официальная цена не кажется неразумной, поэтому вы не ощущаете себя обманутым. Однако иногда абсолютно ничем не обоснованные цифры могут стать для нас «якорями» и заставить принимать нерациональные решения. Амос Тверски и Даниэль Канеман (1974) провели эксперимент: они вращали барабан с нанесенными на него цифрами от 0 до 100, барабан останавливался только на цифрах 10 или 65, но участники эксперимента этого не знали. Для каждого из них вращали барабан, ждали, пока он остановится, и спрашивали, было ли количество африканских стран среди стран, входящих в ООН, выше или ниже этого значения (это этап «якорения»). Затем участников просили оценить процентное соотношение. Те из них, у кого барабан остановился на 10, оценивали примерное соотношение африканских стран в ООН как 25%, тогда как участники, у которых барабан остановился на 65, называли примерное соотношение 45%, — разница в 20% из-за, казалось бы, «случайного» ничего не значащего поворота барабана.

Мы устаем и начинаем испытывать чувство голода

На наши решения влияют такие внутренние факторы, как чувство голода, настроение, уровень энергии. В 2011 году был проведен интереснейший анализ постановлений восьми израильских судей[1]. Данцингер и др. изучили 1112 постановлений суда, выне-

[1] Danzinger S., Levav J. and Avnaim-Pesso L. Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. 108 (2011): 6889–6892.

сенных в течение 50 дней за период десять месяцев. Кроме того, ученые также отслеживали, когда судьи делали перерыв на легкий перекус до обеда (в среднем на 40 минут) и перерыв на обед (около часа). Изначальная предпосылка состояла в том, что самое простое решение — отказать в условно-досрочном освобождении, а самое сложное решение — разрешить его. Во втором случае принятие решения занимало больше времени (пять минут против семи, соответственно) и постановление было длиннее (47 слов против 90). Процент положительных решений (разрешающих условно-досрочном освобождение) начинался с 65% в начале дня и снижался почти до 0% ко времени первого перерыва. После перерыва он поднимался до 65% и постепенно снижался до 0% вплоть до перерыва на обед. Догадываетесь, что происходило после обеда? Процент положительных вердиктов подскакивал до 65% и постепенно снижался до конца рабочего дня. (Эти результаты нельзя было объяснить такими факторами, как расовая принадлежность, тяжесть преступления, срок заключения и другими.) Авторы не могли контролировать, был ли причиной сам факт перерыва или повышение уровня глюкозы в крови после приема пищи, но было очевидно, что внутренние факторы влияют на процесс принятия решения. По словам авторов исследования, «сатира по поводу того, что справедливость зависит от того, что судья ел на завтрак, может относиться к тому, как люди принимают решения в целом».

Я выделил несколько когнитивных искажений, которым мы подвержены. На самом деле их гораздо больше.

Перечислим важные искажения, способные негативно повлиять на наши суждения.

«Ошибка выжившего»

Мы считаем репрезентативными те данные, которые подтверждают успех какого-либо предприятия. Если почитать технологические блоги, такие как Techcrunch, Re/Code или O'Reilly Radar, на вас обрушится лавина историй об успешных стартапах, владельцы которых их запустили, привлекли финансирование и вышли из бизнеса. Начинает казаться, что любой стартап обречен на успех. Но в этих блогах не пишут о том, что подавляющему большинству стартапов не удается выйти на этап привлечения инвестиций, и даже среди тех, кому это удается, 97% или около того не доживают до этапа выхода. Нам становится известно только о тех, у кого это получилось.

Предвзятость подтверждения

Учитывая, что мы «с трудом отказываемся от устаревшей информации», одно из когнитивных искажений связано с тем, что человек ищет или предпочитает выбирать данные, подтверждающие то, что он уже знает. Эйнштейн шутил, когда говорил: «Если факты не подтверждают теорию, смените факты», но тем не менее ученые обнаружили, что именно этим может заниматься левое полушарие человеческого мозга (см. основной доклад Шанкара Вендатама[1]).

Эффект новизны

Мы склонны больше вспоминать недавние события и фокусироваться на них[2]. В большинстве случаев это оправданный подход, хотя и не всегда. Предположим, что на фондовом рынке наблюдается стабильная тенденция на понижение. Только то, что вчера акции немного выросли в цене, не означает, что рынок достиг дна. В условиях стохастической и волатильной среды необходимо расширить временной горизонт, чтобы получить представление об общем тренде, поскольку данные, полученные за короткий промежуток времени, — ненадежная информация.

Эффект «свой-чужой»

Когда кто-то сообщает вам информацию, первое, что вы делаете, — оцениваете собеседника: это друг или враг, конкурент или союзник, — а затем решаете, можно ли доверять этой информации. То есть «люди считают, что солидная и благонадежная внешность — это мотивация говорить правду»[3].

КОГДА ИНТУИЦИЯ РАБОТАЕТ?

Разумеется, бывают ситуации, когда стоит довериться интуиции, и она вас не подведет. К числу часто приводимых примеров относят интуицию опытных пожарных, которые чувствуют, когда находиться

[1] URL: https://www.youtube.com/watch?v=7mpe6luA5Os.

[2] Я обратил внимание, что, когда радиостанции составляют рейтинги любимых песен слушателей «всех времен», в топ-20 преимущественно входят песни, популярные в течение последнего года. Это «эффект новизны» в действии. Он работает и при совершении покупок: если последний опыт совершения покупки был негативным, он перекроет позитивное впечатление, которое сложилось от нескольких предыдущих покупок. Успех определяется лишь тем, насколько успешным был последний раз.

[3] Fiske S. T. and Dupree C. Gaining trust as well as respect in communicating to motivated audiences about science topics, PNAS 111, no. 4 (2014): 13593–13597. URL: http://www.pnas.org/content/111/Supplement_4/13593.full.

в охваченном огнем здании уже опасно, и выводят оттуда свою команду; или опытных медицинских сестер из отделения детской реанимации, которые еще до консультации с врачами и до результатов клинических тестов могут сказать, что у младенца жар или какие-то осложнения; или шахматных гроссмейстеров, способных предугадать игровую стратегию оппонента и оценить, казалось бы, невероятное количество ходов. Подобного рода интуиция может развиться только в условиях, когда «подсказки» и сигналы надежные и постоянные. То есть это возможно, например, в больничном отделении, где пациент проводит несколько дней или недель, взаимодействуя с одним и тем же медицинским персоналом, но это не сработает в условиях быстро меняющейся среды, например на фондовой бирже.

Чтобы развить такую интуицию, потребуется немало времени. Хотя сейчас есть все основания сомневаться в правиле «10 тыс. часов»[1], справедливо, что на определенном уровне практика имеет очень важное значение. У немногих руководителей бывает достаточно времени для работы с узкой и постоянной темой, чтобы стать в ней настоящим экспертом.

В среднем человек меняет место работы от пяти до семи раз (хотя точная цифра, конечно, неизвестна), к тому же у него могут часто меняться должности и профессиональные области внутри компании. Прошли те времена, когда человек мог проработать на одном рабочем месте всю жизнь. Иными словами, мне кажется, что, с точки зрения профессионального опыта, сегодня мы гораздо чаще начинаем всё с нуля.

Интуиция может быть весьма ценным качеством, если используется для проверки фактов. Если данные не соответствуют ожиданиям, это может быть сигналом о необходимости еще раз проверить данные. В главе 2 я уже упоминал о том, что прогнозирование вероятных значений или данных может стать частью проверки качества данных. В отчете Decisive Action говорится: «Интуитивное ощущение может стать предупреждением: на этапе сбора данных или анализа было сделано что-то неправильное. Это позволит руководителю проверить достоверность данных, на которых основываются его решения».

Я был рад услышать ответ на следующий вопрос: «Что бы вы сделали, если бы при принятии решения имеющиеся у вас данные противоречили вашей интуиции?» 57% респондентов сказали, что они провели бы повторный анализ данных, 30% респондентов собрали бы

[1] Macnamara B. N., Hambrick D. Z. and Oswald F. L. Deliberate practice and performance in music, games, sports, education, and professions: a meta analysis, Psychological Science 25 (2014): 1608–1618.

дополнительные данные. Только 10% респондентов продолжили бы с имеющимися данными (рис. 9.6).

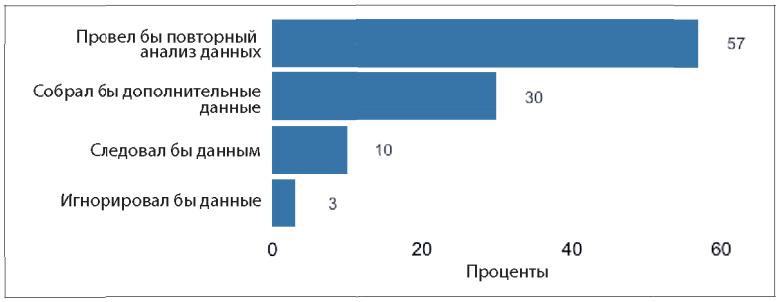


Рис. 9.6. Что бы вы сделали, если бы при принятии решения имеющиеся у вас данные противоречили вашей интуиции?

Источник: отчет Decisive Action

Решения

У вас еще не возникло ощущения безнадежности? Получившаяся картина выглядит довольно уныло. Тогда давайте сменим тон и переключимся на потенциальные решения. Что можно предпринять, чтобы стимулировать процесс принятия решений на основе данных?

В этом разделе я буду оперировать терминами в рамках поведенческой модели Фогга[1]. Если человеческий мозг — источник стольких проблем с принятием решений на уровне интуиции, давайте покопаемся в собственной голове, чтобы понять, как мы можем мотивировать поведение и принимать решения.

Следователи по уголовным делам часто фокусируются на том, были ли у подозреваемого мотив, способ и возможность совершения преступления. При отсутствии хотя бы одного из этих трех компонентов маловероятно, что подозреваемый будет осужден. Поведенческая модель Фогга чем-то напоминает эту триаду. В рамках этой модели формулируется набор условий для выполнения какого-либо действия и предполагается следующее:

— человек должен быть достаточно мотивирован;
— человек должен обладать возможностью выполнить действие;
— на человека должен воздействовать стимул, побуждающий его выполнить действие.

[1] URL: http://bjfogg.com/fbm_files/page4_1.pdf.

Вопрос в том, как создать условия для того, чтобы решения принимались на основе данных, а не на основе интуиции. Давайте изучим этот вопрос с позиции поведенческой модели Фогга.

МОТИВАЦИЯ

Первое условие — наличие мотивации. Что может повысить мотивацию более активно опираться на данные или хотя бы улучшить процесс принятия решений (что предположительно будет включать ориентацию на использование данных)?

Фогг выделяет три типа мотивирующих факторов.

Удовольствие/боль

Примитивный мотиватор немедленного действия.

Надежда/страх

Мотиватор, требующий больше времени.

Социальное принятие/отторжение

По Фоггу, Facebook обладает силой мотивировать своих пользователей и таким образом оказывать на них влияние именно благодаря этому фактору.

Три мотивирующих фактора Фогга можно переложить на реалии бизнес-среды, и мы получим гордость (которая стимулирует сотрудников хорошо выполнять работу ради собственного чувства удовлетворения), удовольствие от признания, похвалу, продвижение за качественное выполнение работы или, наоборот, страх наказания за плохо выполненную работу.

Я наивно полагал, что деньги тоже мотивирующий фактор, особенно в бизнес-среде, где бонусы по итогам года привязаны к показателям эффективности компании. Удивительно, но при решении сложных задач или задач, требующих нестандартного подхода, деньги не только оказались плохим мотиватором, но и *ухудшили* эффективность деятельности[1].

Стимулы и подотчетность

Ранее я уже упоминал об отсутствии подотчетности. Эту ситуацию нужно исправлять. Один из способов, конечно, привязать результаты

[1] URL: https://www.youtube.com/watch?v=u6XAPnuFjJc.

деятельности к количественным показателям, таким как уровень продаж, количество подписок или показатель выручки. Можно сфокусироваться на показателе ROI или общем влиянии на бизнес, хотя чаще всего руководители и так ориентируются именно на эти показатели. Если сотрудник принял неэффективное решение, это должно отражаться в показателях. Разрабатывайте стимулы, чтобы поощрять необходимое вам поведение и развивать корпоративную культуру.

Наличие доказательств

Вместо того чтобы полагаться на шестое чувство, развивайте у себя в компании такую корпоративную культуру, в которой идеи подвергаются сомнениям, пока не будут получены достоверные данные, например результаты A/B тестов, доказательства концепции или результаты моделирования.

Прозрачность

Стимулируйте развитие более открытой и прозрачной корпоративной культуры, чтобы было очевидно, кто и какие решения принимает, а также к каким результатам это приводит. Повышая прозрачность самих решений и результатов этих решений с помощью презентаций, отчетов или дашбордов, вы запускаете мотивирующий фактор социального принятия.

ВОЗМОЖНОСТЬ ВЫПОЛНИТЬ ЗАДАЧУ

По Фоггу, можно выделить шесть аспектов, влияющих на возможность человека выполнить задачу.

Время

Выше вероятность, что человек выполнит краткосрочную задачу по сравнению с долгосрочной.

Деньги

Выше вероятность, что человек выполнит задачу, не требующую серьезных финансовых затрат, чем дорогостоящую задачу.

Физические усилия

Выше вероятность, что человек выполнит задачу, требующую меньше физических усилий.

Умственные усилия

Выше вероятность, что человек выполнит задачу, не требующую серьезных умственных усилий.

Отклонение от социальных норм

Выше вероятность, что человек выполнит задачу, которая является социально приемлемой.

Рутинность

Выше вероятность, что человек выполнит рутинную задачу, чем неординарную.

Руководствуясь этими принципами, относительно просто понять, как можно снизить барьеры для принятия хороших решений. В последующем обсуждении я с помощью скобок буду выделять шесть перечисленных возможностей.

Привяжите действия к результатам

Аналитики могут облегчить процесс принятия решений (умственные усилия) для руководителей и снизить время принятия решений (время), если подберут правильную форму для презентации своих выводов и рекомендаций, отразят, почему это важно, и сфокусируются на влиянии. Да, следует представлять доказательства и рекомендации в наиболее доступной форме, чтобы для их понимания требовалось минимальное усилие. Мне нравится форма презентации, которую предложила Трейси Эллисон Олтмен. Эта форма представлена на рис. 9.7 (остальная работа Атмен тоже достойна внимания) и выделяет взаимосвязь между действием и результатом: если вы сделаете X, то случится Y. Кроме того, она подтверждает рекомендации, следующие далее. Это и есть сделка: «купите» эти рекомендации в силу объективных причин.

По результатам опроса компании Accenture[1], 58% руководителей считают, что самое сложное — увидеть результаты от работы с данными: «Установление взаимосвязи между сбором данных и проведением анализа и действиями и результатами, спрогнозированными аналитиками, для многих оказывается более сложной задачей, чем сбор или интерпретация данных». Более того, как оказалось, только 39% руководителей считают данные, которые приводят аналитики, «релевант-

[1] URL: https://www.accenture.com/us-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Technology_6/Accenture-Analytics-In-Action-Survey.pdf.

ными для бизнес-стратегии». Именно здесь каждый специалист, работающий в компании с данными, должен сыграть свою роль. Помогите включить аналитику в бизнес-процесс, сделать ее более прозрачной и понятной, более постоянной с адекватными данными и показателями. Выражайте свое несогласие, если это необходимо, но будьте готовы объективно доказать свою точку зрения.

Рис. 9.7. Привяжите действия к результатам. Укажите действие с привязкой к конкретному результату, а ниже представьте причинно-следственное доказательство

Источник: https://www.uglyresearch.com/datatodecision.php. Воспроизводится с разрешения

Сотрудничество и согласие

В главе 5 я уже рассказывал, как Нейту Сильверу удалось предсказать результаты выборов в Сенат и победителей в 49 штатах из 50 в ходе предвыборной кампании 2008 года. Он сделал это, после того как ученые мужи высмеяли его, утверждая, что, благодаря своему огромному опыту в области политологии, они всё знают лучше него. Однако построение статистических моделей на основе совокупности разных опросов и мнений (а также с использованием самых последних данных, которые только можно было получить) позволило Сильверу сделать прогноз с высоким уровнем точности, в котором были усреднены различные

ошибки. Как отметил Ларри Кили из Doblin Group, «хорошие идеи могут прийти от кого угодно» (цит. по книге Кевина Келли New Rules for the New Economy (Penguin Books)). В данном случае «кто угодно» — это электорат, мнение которого отражено в агрегированных данных.

Если решение сложное или непопулярное, одним из вариантов становится достижение согласия (отклонение от социальных норм). Это даст право голоса всем заинтересованным сторонам и повысит шансы на успех. «Важно, чтобы каждый ощущал себя частью процесса. Нет никакой пользы в эффективном решении, если его никто не поддерживает», — отмечает Робин Тай, исполнительный директор Ernst and Young.

В современной реальности это означает, что все сотрудники должны понимать цели, характер собираемых данных, показатели и то, как руководитель интерпретирует информацию при принятии решений. Обеспечьте сотрудникам возможность выразить свою точку зрения, если она отличается от вашей, и участвовать в процессе. При этом проанализируйте другие варианты, которые, возможно, упустил руководитель. В качестве подсказки можно воспользоваться акронимом DECIDE.

— Определите проблему (**D**efine).
— Установите критерии (**E**stablish).
— Рассмотрите все альтернативы (**C**onsider).
— Выделите лучшую (**I**dentify).
— Разработайте план действий и начните его воплощать (**D**evelop).
— Оцените решение и при необходимости дайте обратную связь (**E**valuate).

Иными словами, убедитесь, что все участники процесса согласны с этим шагами.

Конечно, у такого подхода есть свои минусы. Если в процессе принятия решения задействовано слишком много людей, это может привести к эффекту коллективного мышления, а также к размытию ответственности, что может существенно замедлить процесс принятия решения или повысить вероятность появления противоречащих друг другу позиций, что способно спровоцировать споры и разногласия. Опять-таки, здесь необходимо найти золотую середину, то, что подтверждено данными (рис. 9.8).

Интересно, что, согласно данным отчета Decisive Action,

> …в то время как топ-менеджмент компании и руководители подразделений чаще всего опираются в своих решениях на данные, вице-президенты и старшие вице-президенты

(или сотрудники на эквивалентных должностях), по их собственной оценке, более склонны к совместному принятию решений. Это может быть признаком того, что руководителям этого уровня требуется заручиться более широкой поддержкой своей инициативы, что перестает быть актуальным для руководителей высшего звена.

Рис. 9.8. Распределение ответов на вопрос «К мнению скольких сотрудников вы прислушиваетесь, принимая решения в вашей компании?»

Источник: отчет Decisive Action: how businesses make decisions and how they could do it better

Обучение

Повышение статистической грамотности людей, принимающих решения, — очевидный шаг для улучшения возможности предпринимать действия (умственные усилия). Конечно, проведение статистического анализа — обязанность аналитика, так что вряд ли всем руководителям нужно уметь строить сложные регрессионные модели или понимать математические основы EM-алгоритма или метода опорных векторов.

Вместо этого я рекомендовал бы сосредоточиться на принципах формирования выборок и разработки экспериментов, чтобы те, кто принимает решения, могли оценить качество собранных данных и достоверность результатов тестирования, какие факторы могут повлиять на объективность данных и так далее. Кроме того, я рекомендовал бы провести обзор показателей с возможными отклонениями, такими как предел погрешности и стандартное отклонение, которые отражают воспроизводимость и уверенность в итоговых совокупных значениях.

Внимание: при попытках провести подобного рода обучение вы можете натолкнуться на сопротивление, так что, возможно, вам придется заручиться поддержкой руководителей самого высокого уровня

(как это было у нас в компании Warby Parker), чтобы убедить всех заинтересованных людей пройти курс повышения квалификации, пусть даже продолжительностью всего час.

Постоянство

Выполнение задач можно сократить по времени (время) и сделать проще (умственные усилия) благодаря единообразию в презентации данных. Это не означает, что все отчеты должны выглядеть одинаково, тем не менее форма еженедельного отчета или дашборда не должна меняться со временем. Кроме того, по возможности команды должны получать одни и те же показатели.

Например, в корпорации Procter & Gamble, где дашбордами пользуются 50 тыс. сотрудников, унификация данных для всех пользователей — необходимость. На интерактивной карте, отражающей долю рынка корпорации, зеленый цвет всегда обозначает «выше рыночной доли», а красный — «ниже рыночной доли». Не стоит без необходимости смешивать показатели. Кроме того, в корпорации разработаны модели достаточности (business sufficiency models[1]), которые определяют, какие данные необходимы для работы в определенной профессиональной области. Это означает, по Томасу Дэвенпорту, что «если вас, например, интересуют вопросы цепочки поставок, модель достаточности определяет основные переменные, как они должны быть представлены визуально и (в некоторых случаях) взаимосвязи между переменными и прогнозами на основе этих взаимосвязей».

ПОБУЖДАЮЩИЕ СТИМУЛЫ

Из трех факторов по модели Фогга наличие побуждающего стимула, вероятно, наименее важно, по крайней мере, в контексте принятия деловых решений. Я говорю это, потому что решения в бизнесе обычно принимаются в более широком контексте целей, основных показателей эффективности, стратегии и совместной командной работы, где обычно присутствует реальный или установленный срок выполнения задачи. То есть если кто-то не спрашивает о решении или не ждет его, очевидно, что в процессе что-то явно не так или это не слишком важно. Конечно, сложное решение всегда можно попробовать отложить под реальным или вымышленным предлогом нехватки данных. С этим

[1] URL: https://hbr.org/2013/04/how-p-and-g-presents-data.

можно бороться, если установить четкий, прозрачный график проекта и распределить зоны ответственности.

Один из примеров, когда действительно есть необходимость в побуждающем стимуле, — автоматический процесс, которым «управляют» статистические модели с принципами машинного обучения. Подобные модели устаревают. Внутренние предположения, на основе которых они строились, теряют актуальность, например поведение потребителей или сотрудников (как один из движущих факторов) может измениться. Таким образом, требуется регулярно проверять эффективность этих моделей, проверять предположения и по мере необходимости вносить коррективы. При этом, когда во главу угла ставится алгоритм, управляющий процессом, люди становятся более пассивными и теряют бдительность: проявляется так называемый эффект автоматизации. Для преодоления этого эффекта нужно установить четкий график и обязанность поддерживать актуальность модели.

Заключение

Процесс принятия решений бывает непростым. Мы подвержены воздействию самых разных факторов, способных повлиять на объективность принимаемых решений. Это в том числе когнитивные искажения, проблемы с данными и корпоративной культурой компании. Помешать принимать объективные решения может предвзятое мнение или раздутое эго.

Интуиция должна стать частью процесса принятия решений на основе данных. Без нее не обойтись. В заключении своей книги Dataclysm Кристиан Раддер признает: «За каждой цифрой стоит человек, принимающий решение: что анализировать, что исключить из процесса анализа, в какую рамку поместить ту картину, которую рисуют данные. Сделать заявление, построить простейший график — означает сделать выбор, и при этом несовершенство человеческой натуры непременно даст о себе знать».

Скотт Беркен также отмечает: «Когда кто-то говорит “данные показывают”, он притворяется, что существует единственная интерпретация этих данных, но это далеко не так. Подобное ложное убеждение мешает задавать важные вопросы, например “Можно ли на основании этих же данных выстроить альтернативную и в равной степени убедительную гипотезу, ведущую к другому заключению?”»

Основное в этом процессе — начать с правильных вопросов и сконцентрироваться на вопросе и решении[1], а не на данных. Когда вы четко и недвусмысленно формулируете свою цель, у вас увеличивается вероятность правильно определить, на какие вопросы нужно ответить и, следовательно, какие данные собрать, какие тесты провести, какие показатели продвигать. Таким образом, у вас увеличивается вероятность, что полученные результаты будут соответствовать вашим показателям и целям, а принимать решения вам будет проще.

Тем не менее вы обязательно должны использовать имеющиеся в вашем распоряжении релевантные данные. Не стоит полагаться исключительно на интуицию, она слишком часто подводит. Что еще важнее — не сдавайтесь на милость HiPPO. Если вы вынуждены принять решение, идущее вразрез с данными, отдавайте себе отчет, когда и почему вы это делаете и ради какой цели, например для реализации долгосрочной стратегии (как в примере с Amazon из главы 8).

Рис. 9.9. Что из перечисленного, по вашему мнению, больше всего способствовало бы улучшению процесса принятия решений в вашей компании?

Источник: на основе диаграммы 7 из отчета Decisive Action: How businesses make decisions and how they could do it better аналитического подразделения журнала Economist

[1] URL: https://www.uglyresearch.com/datatodecision.php.

Мы рассмотрели ряд вопросов, важных на этапе принятия решения, включая данные и когнитивные аспекты. Какие из них руководители считают наиболее важными или наиболее легкодостижимыми? Двумя самыми популярными ответами были улучшение способности анализировать данные и повышение подотчетности при принятии решений (рис. 9.9). Реализовать оба этих аспекта относительно просто. Тем не менее достижимы все перечисленные факторы, хотя это и требует поддержки всех сотрудников — от специалистов по сбору данных до топ-менеджмента компании. Добиться этого возможно только в условиях соответствующей корпоративной культуры и при наличии мотивированных сотрудников с правильными стимулами. Как отметил один из комментаторов, «будучи аналитиком, я могу утверждать, что в очень многих компаниях представлять данные, противоречащие точке зрения или намерениям HiPPO, — прямой путь к увольнению и попаданию в черный список»[1]. В компании с управлением на основе данных это неприемлемо. Таким образом, мы переходим к вопросу корпоративной культуры, что и будет темой следующей главы.

[1] URL: https://plus.google.com/+JonathanRosenberg/posts/DaUY9tT8Ev6.

ГЛАВА 10

Корпоративная культура на основе данных

Самая большая проблема, с которой сталкиваются компании, пытающиеся внедрять инновации и трансформироваться, — корпоративная культура по типу «мы всегда так делали».

Габи Боко[1]

Корпоративная культура на основе данных — это не только применение новейших технологий, это изменение традиционной корпоративной культуры так, чтобы компания, команды в ней и каждый сотрудник стремились делать что-то отличное, потому что располагают для этого необходимыми данными.

Сатья Наделла[2]

Важность корпоративной культуры — тема, которая красной нитью проходит через всю книгу. По мере того как данные продвигаются по аналитической цепочке ценности, можно выделить ряд контактных точек: некоторые из них связаны с людьми, некоторые — с технологиями, но все они зависят от преобладающей в компании корпоративной культуры. Корпоративная культура определяет, кто имеет доступ к данным, какие данные можно распространять, какие вложения будут сделаны в развитие сотрудников и в инструменты. Более того, как я уже отмечал в предыдущей главе, корпоративная культура определяет, HiPPO или данные будут влиять на последнее звено в цепочке.

В этой главе мы подробнее остановимся на всех этих аспектах и рассмотрим их в совокупности, чтобы представить единую и полную картину

[1] Economist Intelligence Unit. The Virtuous Circle of Data: Engaging employees in data and transforming your business (London: Economist Intelligence Unit, 2015). URL: http://live.wavecast.co/virtuouscircleofdata/.

[2] Nadella S. A data culture for everyone, The Official Microsoft Blog, April 15, 2014. URL: https://blogs.microsoft.com/blog/2014/04/15/a-data-culture-for-everyone/#sm.00000q4vufg9naev6waguvc6wipz7.

идеальной компании с управлением на основе данных. Мы начнем с основ работы с данными: с доступа к данным, обмена ими и широкого обучения, как их использовать. Затем мы перейдем к обсуждению корпоративной культуры, где сначала ставятся цели, разрабатываются критерии успеха, показатели и схема эксперимента, а после существует возможность обсуждения результатов эксперимента, их интерпретации и анализа. За этим последует обсуждение итераций, обратной связи и обучения. Завершим мы обсуждением того, как противодействовать HiPPO и как организовать управление на основе данных «сверху вниз».

В некотором смысле перечисленные темы, или критерии, можно считать списком основных ингредиентов. Представьте, сколько разных тортов и пирожных можно испечь, имея муку, яйца, масло и сахар. Итоговый результат будет зависеть от качества продуктов, их пропорции и сочетания. Точно так же и с компаниями с управлением на основе данных. Они могут быть самыми разными. Вы должны выбрать ту форму, которая подходит для вас, учитывая вашу стартовую площадку, область деятельности, размер и зрелость компании. Более того, не стоит ожидать, что вы достигнете волшебной точки равновесия, — ваша компания будет постоянно меняться. Вы должны инвестировать в развитие, экспериментировать и запастись терпением.

Открытость и доверие

> *Руководители должны думать о том, как поощрять сотрудников, распространяющих данные, как стимулировать сотрудников и отделы, развивающие и поддерживающие открытые, точные и доступные для использования данные и аналитику.*
>
> Дженнифер Кобб[1]

В компании с управлением на основе данных, как правило, бывает обеспечен широкий доступ к информации. В том числе доступ к данным имеют сотрудники вне аналитического подразделения, к которым относятся все остальные бизнес-единицы, команды и сотрудники. Давайте рассмотрим этот аспект.

В главе 3 мы приводили пример покупки набора садовой мебели Белиндой Смит и то, как использование данных из разных источников

[1] Cobb J. Data Tip #2 — Build a Data-Driven Culture, Captricity Blog, October 30, 2013. URL: http://captricity.com/blog/data-tip-2-build-a-data-driven-culture/.

расширило контекст и улучшило понимание намерений, мотивации и интересов покупателя. Лучше понимая контекст, компания способна обеспечить обслуживание клиентов на более высоком уровне, а также предложить именно те товары, которые требуются пользователю.

Давайте пока оставим в стороне такие внешние источники данных, как Бюро переписи населения и единую базу данных недвижимости (MLS), и остановимся на некоторых внутренних контактных точках клиента и онлайн-продавца:

— история посещений на сайте компании;

— история покупок, возвратов и обменов;

— взаимодействие с сотрудниками службы по работе с клиентами посредством электронной почты, чата, телефона;

— взаимодействие с брендом через социальные сети;

— данные социальных сетей, например через программу «приведи друга»;

— демонстрация бренда через ретаргетинг.

Несложно понять, что обычно этими источниками данных управляют разные команды или бизнес-подразделения. Для максимально эффективного использования данных в компании эти данные необходимо собрать вместе, чтобы получить более полный контекст. И здесь вступает в действие корпоративная культура.

Следует четко дать понять, что данные — это не собственность конкретного подразделения, они принадлежат всей компании. Руководители направления по работе с данными (о них мы поговорим далее) должны рассказывать о преимуществах информационной открытости внутри компании. Однако, если это не сработает, у компании должны быть правильные стимулы, чтобы преодолеть разобщенность и наладить обмен данными.

Конечно, проводить подобную политику следует в соответствии со всеми нормами и правилами и не в ущерб конфиденциальности и безопасности. Эти опасения не беспочвенны. Треть респондентов, участвовавших в опросе[1] 530 руководителей, который провело аналитическое подразделение журнала Economist, отметили, что их компании «не удается воплотить корпоративную культуру на основе данных частично из-

[1] URL: https://www.tableau.com/economist-fostering-data-driven-culture.

за вопросов конфиденциальности и безопасности, которые возникают при обмене данными».

По причине этого обоснованного беспокойства, но также по инерции, режим, в котором руководители бизнеса действуют по умолчанию, — это режим накопления данных. С этим нужно активно бороться. В том же опросе руководителей спросили, какие стратегии они считают успешными для продвижения корпоративной культуры на основе данных. В результате в качестве одной из главных стратегий был указан пункт «Продвижение способов обмена информацией» (он совсем немного уступил пункту «Прямые указания со стороны руководства») (рис. 10.1).

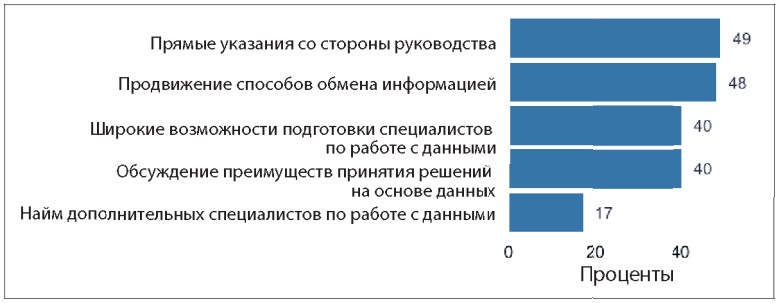


Рис. 10.1. Распределение ответов на вопрос «Какие стратегии доказали свою эффективность в продвижении корпоративной культуры на основе данных в вашей компании?»

Источник: опрос 530 руководителей, проведенный аналитическим подразделением журнала Economist

Для обмена данными требуется определенный уровень доверия. Во-первых, сотрудники должны быть уверены, что этим данным можно доверять, что они надежны и точны. Во-вторых, сотрудники должны быть уверены, что данные будут использованы во благо, а не обернутся против них.

Например, в одной из больниц[1] «врач боялся, что его медицинские записи увидят коллеги, которые могут найти у него ошибку». Люди должны преодолеть подобные страхи и сосредоточиться на повышении качества данных. В-третьих, и это связано со второй темой этого раздела, данные должны предоставляться всем сотрудникам компании.

[1] URL: https://hbr.org/resources/pdfs/tools/HBR_Qlik_Report_May2014.pdf.

Компании с управлением на основе данных отличаются большей открытостью и прозрачностью, данные демократизированы и доступны многим сотрудникам. «У каждого сотрудника компании должен быть доступ к такому количеству данных, которое только возможно на законных основаниях», — утверждают Ди Джей Патиль и Хилари Мейсон[1] (см. также главу 12). Доступ к данным может осуществляться через отчеты и дашборды, но может быть и «активным» за счет использования инструментов бизнес-аналитики и даже необработанных данных. Это также требует значительного доверия. Компания должна быть уверена, что не произойдет утечки информации к конкурентам, что ее данные не будут использовать в различных политических интригах, а станут исключительно способствовать росту и развитию бизнеса.

Если идти дальше, то компания с управлением на основе данных обладает более значительным потенциалом делегировать принятие определенных решений на операционный уровень. Если у большего числа сотрудников есть доступ к нужным им данным, имеются необходимые навыки их анализа и интерпретации, то при достаточном уровне доверия процесс принятия решений можно существенно демократизировать. Например, предположим, что менеджер розничного магазина обладает навыками работы с инструментами бизнес-аналитики, благодаря чему он способен проанализировать уровень продаж единиц складского учета в своем магазине, определить сезонные колебания, принять во внимание местные особенности, такие как климатические условия, качественно прогнозировать тренды и делать заказы на продукцию так, чтобы у него не было дефицита товара, но хранился минимальный запас на складе.

Очевидно, что многие решения, особенно важные или стратегические, все равно будут приниматься на уровне высшего руководства. Тем не менее в большинстве компаний многие решения, особенно касающиеся операционной деятельности, можно делегировать на места, если обеспечен доступ к нужным данным, а также есть необходимые навыки и соответствующий уровень доверия. Можно провести аналогию с нервной системой человека. Большинство решений отправляются в головной мозг на обработку, но если вы наступили, скажем, на кнопку, проявляется спинно-мозговой рефлекс, когда стимул достигает спинного мозга, откуда мышцам поступает команда убрать ногу. «Местной» обработки информации и принятия решения достаточно для разрешения этой проблемы.

[1] URL: http://www.oreilly.com/data/free/data-driven.csp.

Повышение квалификации в области работы с данными

Если организация стремится внедрить подход, ориентированный на данные, стимулировать корпоративную культуру, в которой понимают и ценят данные, тогда отличное понимание данных должно входить в навыки и характеристики всех сотрудников всех уровней, особенно в коммерческой компании.

Отчет компании Accenture[1]

Очевидно, что специалисты по аналитической работе должны пройти обучение по планированию экспериментов, развитию навыков критического мышления, презентации данных, применению инструментов бизнес-аналитики и статистики и так далее. Однако чтобы вся компания стала ориентированной на данные, этот набор навыков, а также подход, опирающийся на факты и доказательства, должен быть внедрен на более широком уровне. Кроме того, руководители и другие лица, ответственные за принятие решений, также должны быть компетентны в области работы с данными. Почему это важно?

— Руководители подписывают счета на приобретение, установку и обеспечение работоспособности новых инструментов бизнес-аналитики или сервисов прогнозного моделирования. Они должны понимать ценность этих инструментов для компании.

— Руководители соглашаются на вре́менные неудобства для рабочего процесса и на снижение эффективности работы, когда специалисты по аналитике уходят на повышение квалификации или осваивают новые инструменты. Иными словами, чтобы согласиться на трудности в переходный период, руководители должны видеть выгоду в долгосрочной перспективе.

— Руководители принимают ключевые стратегические и тактические решения на основе аналитических выводов. Они должны быть в состоянии увидеть недостатки в проведенном анализе и вернуть его на доработку, если анализ выполнен некачественно. Они постоянно должны требовать более глубоких и качественных данных и ожидать от аналитика большего. Кроме того, руководителям приходится представлять аналитические выводы высшему руководству компании, совету директоров или инвесторам. То есть

[1] Accenture Technology Vision 2012. Data Culture. URL: https://www.accenture.com/us-en/new-applied-now.

они должны понимать особенности проведенного анализа, быть уверены в выводах и рекомендациях и быть готовы их отстаивать.

Иными словами, руководитель необязательно должен владеть механизмами сбора, очистки, обработки и агрегирования данных, но у него должно быть понимание, что такое качественный эксперимент, базовое статистическое исследование, а также чем опасно экстраполирование. Например, однажды мне довелось наблюдать, как аналитик представил руководителю результаты анализа, которые мне показались качественно подготовленными и понятными, на что руководитель спросил: «А что такое p-значение[1]?» Конечно, обязанность аналитика — представить результаты анализа в понятном для аудитории формате, но при этом, мне кажется, в компании с управлением на основе данных в зону ответственности руководителя должно входить знакомство с базовой терминологией, показателями и тестами.

Дэвенпорт и др. (Analysts at Work, с. 15) разделяют эту точку зрения:

> По мере того как финансовая и инвестиционная области (а вместе с ними и все остальные отрасли) становятся всё более ориентированными на данные и аналитику, у топ-менеджеров просто не остается другого выхода, кроме как в той или иной степени овладеть навыками аналитической работы. В противном случае они просто не смогут отклонить рискованное предложение какого-нибудь брокера, подвергнув опасности свою компанию и клиентов.

Поддержал это мнение и Брайн д'Алессандро на конференции Strata+Hadoop World[2]:

> Если вы линейный руководитель или топ-менеджер в компании, активно работающей с данными, и если у вас в команде есть специалисты по работе с данными, вы не обязаны знать, как строить прогнозные модели или пользоваться инструментами анализа данных, но определенный уровень компетентности в вопросах статистики у вас должен быть, потому что в один прекрасный день они придут к вам

[1] P-значение — величина, используемая при тестировании статистических гипотез. Наименьшая величина уровня значимости, при которой нулевая гипотеза отвергается для данного значения статистики критерия. *Прим. перев.*

[2] URL: https://conferences.oreilly.com/strata/stratany2014/public/schedule/detail/37642.

> с презентацией в Power Point или отчетом, и именно вы окажетесь тем, кто должен будет критически оценить любой предоставленный анализ.

Итак, что же можно предпринять? Согласно недавнему докладу[1], «компании с управлением на основе данных более активно предлагают своим сотрудникам обучение и поддержку в реализации этого подхода на практике по сравнению с компаниями, где управление на основе данных не применяется (67% против 53%)». В своем выступлении на конференции Strata+Hadoop[2] в 2013 году Кен Рудин описал подход, применяющийся в компании Facebook, — data camp (лагерь по обучению работе с данными). Это две недели интенсивной работы с полным погружением в тему, причем принять участие могут не только аналитики, но и менеджеры проектов, дизайнеры, финансовые специалисты и специалисты по работе с клиентами. Отдельный лагерь проводится для технических специалистов. В первой половине дня участники лагеря в течение трех часов слушают лекции, часть из которых посвящена инструментам работы с данными Facebook. После обеда они работают над выбранными актуальными бизнес-проблемами. Работая на протяжении двух недель с наставником, они учатся исследовать данные, выдвигать гипотезы, задавать правильные бизнес-вопросы, повышают свою квалификацию в вопросах работы с данными. Вот что говорит Рудин:

> Если мы продолжим наше начинание, а я думаю, что у нас все получится, то мы сформируем корпоративную культуру, где каждый будет понимать, что должен использовать данные как часть своей работы. Проводить анализ должен каждый[3].

Конечно, не каждая компания располагает ресурсами, персоналом и стремлением проводить такие программы. Но любая компания может с чего-то начать, к тому же сейчас доступно множество ресурсов. Бесплатные онлайн-курсы по статистике предлагают Coursera, Udacity, Khan Academy и многие другие. Есть отличная литература по теме. Мне нравится бесплатный открытый ресурс OpenIntro Statistics[4]. Однако выбирать литературу или набор обучающих материалов следует так, чтобы они соответствовали уровню аудитории. Главное,

[1] URL: http://live.wavecast.co/virtuouscircleofdata/.

[2] URL: https://www.youtube.com/watch?v=RJFwsZwTBgg.

[3] URL: http://fortune.com/2013/06/13/what-i-learned-at-facebooks-big-data-bootcamp/.

[4] URL: https://www.openintro.org/stat/textbook.php.

начать что-то делать и стимулировать сотрудников — не только из аналитического отдела — развивать навыки работы с данными и инструментами бизнес-аналитики, чтобы они чувствовали себя комфортно в этой теме.

Сначала цели

Алиса: Подскажите, пожалуйста, куда мне отсюда идти?
Чеширский кот: Это зависит от того, куда ты хочешь попасть.
Льюис Кэрролл. «Алиса в Стране чудес»

В сфокусированной компании, независимо от того, осуществляется ли в ней управление на основе данных, есть четкое направление развития и известное всем представление, как должен расти бизнес. Задача руководителя — объединить людей вокруг этого ви́дения и стимулировать их совместную работу для достижения общей цели. В компании с управлением на основе данных эта цель будет более прозрачной, с четко определенными показателями эффективности деятельности и другими связанными показателями, с ясными задачами и текущим положением дел. Эта система показателей должна быть доступна всем сотрудникам компании, чтобы каждый из них понимал, как его действия способствуют достижению главной цели.

Набор основных целей и показателей KPI затем будет спускаться на уровень бизнес-единиц, где в соответствии с ними могут вырабатываться показатели эффективности для этой конкретной бизнес-единицы, которые, в свою очередь, могут стать основой для разработки показателей и целей более низкого уровня. В какой-то момент вы дойдете до индивидуальных проектов, то есть примерных единиц «работы», требующих постановки конкретной цели и установления критериев успеха. При этом заранее определять критерии успеха следует не только при проведении A/B-тестирования (глава 8), а в любом аналитическом проекте. При работе с данными всегда есть возможность вернуться и выбрать тот набор данных, который поддерживает нужное направление и в той или иной степени демонстрирует положительный показатель ROI. Именно поэтому в интересах объективности в компании с управлением на основе данных должна сложиться такая культура, где сначала формируют цели и показатели, и данные под них не подтягивают[1].

[1] Подробнее о ведении проектов по работе с данными см. Max Shron's Thinking with Data (O'Reilly) и Judah Phillips's Building a Digital Analytics Organization (Pearson FT Press).

В случаях, когда решение по поводу следующего шага приходится принимать на основе нескольких переменных, причем некоторые из них отражают плюсы решения, а некоторые — минусы, постарайтесь определить относительный вес или ранжировать эти переменные *до начала* процесса сбора данных. То есть если в рамках подхода требуется построить матрицу взвешенного решения, постарайтесь как можно раньше оценить «удельный вес» всех факторов. Предположим, вам нужно выбрать одного поставщика услуги из нескольких, и вы руководствуетесь такими факторами, как цена, объем и качество. Скорее всего, цена и качество в данном случае образуют негативную корреляцию. После этого достаточно просто обосновать относительный вес факторов, в результате чего кто-то из поставщиков выбьется в лидеры. Благодаря определению относительной важности каждой из трех переменных до сбора данных, вы четко даете понять, что важно для компании, и снижаете возможность подтасовать результаты или выбрать только те данные, которые поддерживают нужное решение.

Задавайте вопросы

«У вас есть данные, подтверждающие это?» — никто не должен бояться задавать этот вопрос (и все должны быть готовы на него ответить).

Джули Арсенолт[1]

В главе 8 я высказал мнение, что когда в компании начинают активно применять тестирование и эксперименты, то фокус обсуждений смещается с мнений на гипотезы, которые могут подвергнуться объективной проверке. Поскольку это всего лишь гипотезы, а не демонстрация власти или опыта, кто угодно в компании может их высказывать. Это не означает, что каждый будет бросаться тестировать любую безумную идею, которая могла у него возникнуть. В расчет принимается множество факторов, таких как брендинг, юзабилити, стоимость разработки и риски. Тем не менее чем шире круг лиц, предлагающих идеи, тем разнообразнее набор этих идей. (Как вы помните, «хорошие идеи могут появиться у любого» и «дайте право голоса молодым специалистам».)

Помимо того, чтобы дать каждому право голоса, в компании с управлением на основе данных должна поощряться атмосфера здоровой

[1] Arsenault J. How to Create a Data-driven Culture. PagerDuty, October 2, 2014. URL: http://fortune.com/2013/06/13/what-i-learned-at-facebooks-big-data-bootcamp/.

любознательности. Нужно стимулировать конструктивные обсуждения, в ходе которых участники запрашивают дополнительную информацию, подвергают сомнениям предположения, обсуждают результаты тестирования или необходимость проведения дополнительных тестов. Презентации и анализы должны снабжаться ссылками на первоначальные данные. Честное и открытое обсуждение возможных проблем с опытным образцом или интерпретацией, а также предложение улучшений пойдет только на пользу развитию бизнеса. Главное, сохранять нейтральный тон обсуждения: мы обсуждаем данные, а не людей.

Наглядный пример подобного подхода — наука. Одна из основных задач классического западного обучения — сделать молодых ученых максимально *объективными*. Частью этой культуры стали активные попытки деперсонализировать их работу. Если раньше научные статьи писались в активном залоге, то примерно с 1920-х годов окончательно оформилась тенденция использовать пассивный залог[1]. Эта тенденция продолжается по сей день.

Конечно, читать статьи в пассивном залоге менее интересно, но это подчеркивает идею о том, что результаты касаются проводимого эксперимента или самих данных, а не людей, которые этот эксперимент проводят.

В компании с управлением на основе данных должно стимулироваться такое же объективное отношение. Если A/B-тестирование сайта показывает, что более крупная кнопка оформления и оплаты заказа не влияет на показатель выручки или коэффициент конверсии по сравнению с той маленькой кнопкой, которая есть сейчас, значит, так тому и быть. В этом никто не виноват. Это объективная реальность. Порадуйтесь, что вы получили новые ценные данные. (Вы можете использовать это свободное место на экране для чего-то другого.)

Майкл Немшофф высказался еще более определенно:

> Поощряйте несогласие. Нет ничего плохого в том, чтобы поставить под сомнение сложившийся ход вещей, если это подкреплено данными. Не во всех компаниях топ-менеджмент позволяет высказывать необычные и отличающиеся предположения. Если приоритет для вас — создание компании с управлением на основе данных, то вы должны принять наличие определенного уровня несогласия.

[1] Например, активный залог (фокус на субъекте действия): «Мы применили удобрения для растений», — или пассивный залог (фокус на объекте): «Растения были удобрены».

> В некоторых случаях несогласие стоит даже награждать. С разрешения топ-менеджмента компании нужно учить сотрудников уходить с проторенных троп. Новые идеи — подтвержденные данными — отличная стартовая площадка для положительных инноваций[1].

Итерации и обучение

Ошибки — это порталы открытий.
Джеймс Джойс

В предыдущей главе мы говорили о том, что недостаток подотчетности был назван одной из основных проблем в отношении людей, принимающих решения. Кто-то должен «вести счет», не только чтобы люди, принимающие решения, за них отвечали, но и чтобы у компании была возможность учиться и расти. Например, предпринимая определенные действия на перспективу, такие как построение прогнозных моделей, важно не забывать о петле обратной связи, в рамках которой вы проводите регулярный обзор результатов, изучаете отдельные случаи (так называемый анализ ошибок), выясняете, где вы могли бы действовать эффективнее.

Какое-то время я был специалистом по работе с данными в компании One Kings Lane — интернет-магазине по флеш-распродажам товаров для дома. Каждое утро мы предлагали пользователям 4 тыс. наименований товаров, 60% из которых не выставлялись ранее. (Все эти предметы были в ограниченном количестве, и мы продавали их в течение трех дней или пока товар не закончится, в зависимости от того, что происходило быстрее.) Мы с коллегами строили наборы моделей, прогнозирующие, сколько товаров будет распродано к концу одного дня и к концу трех дней. У нас был дашборд, отражавший наши ошибки прогнозирования. Каждое утро мы проводили около часа, изучая и анализируя эти ошибки. Почему нам не удалось правильно спрогнозировать продажи этих ковриков? Действительно ли пользователи случайным образом выбирают между очень похожими товарами? Наша повседневная рутина превращалась в увлекательное занятие, частично потому, что мы относились к этому как к дружескому соревнованию. Мы обменивались идеями, начинали лучше понимать данные, и качество наших моделей неизменно росло. Причина была в постоянных

[1] URL: https://pages.questexweb.com/FierceTechExec-Pub-Signup_FierceTechExec-Signup-Offer.html.

итерациях и обратной связи, в непрерывном анализе пограничных случаев, в попытках их понять и улучшить общее качество.

То же верно и в отношении тестирования и экспериментов. Как уже говорилось в главах 8 и 9, интуиция часто нас подводит. Более половины онлайн-экспериментов ни к чему не приводят. Однако это совсем не провал, если вы анализируете причины и учитесь на своих ошибках.

На рис. 10.2 показана общая петля обратной связи. Вы планируете и проводите эксперимент, измеряете результаты, анализируете данные, интерпретируете результаты, делаете выводы, строите гипотезы и планируете новый эксперимент. Достигаете верхней точки и вновь начинаете движение по кругу. Планирование эксперимента — условное название для этого этапа. С таким же успехом его можно назвать построением модели или разработкой PR-кампании. Суть в том, что компания с управлением на основе данных должна извлекать максимальную пользу из *любых* данных, даже если это был «провал», учиться на своих ошибках и *действовать* дальше, продвигая бизнес.

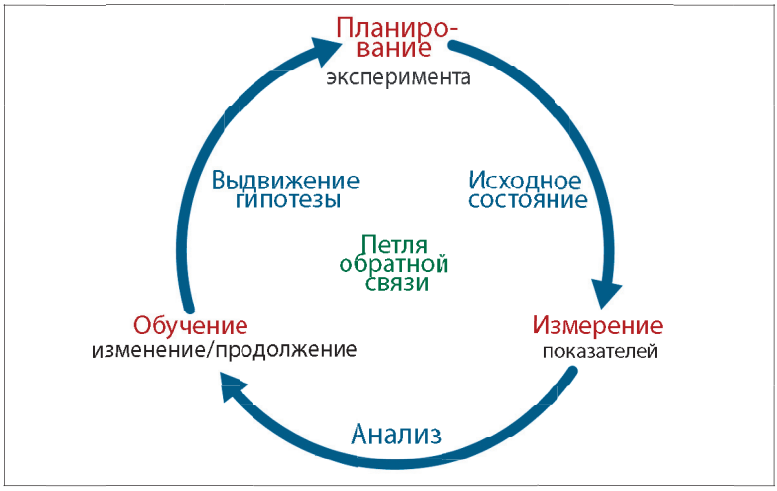


Рис. 10.2. Петля обратной связи: планирование, измерение, выводы и повтор.

Источник: на основе рисунка Эндрю Фрэнсиса Фримена. Воспроизводится с разрешения

Этот аспект должен глубоко укорениться в корпоративной культуре компании. В компании с управлением на основе данных, где все сотрудники наблюдают за данными, любой может выдвинуть гипотезу и большинство сотрудников используют данные в работе, как правило,

наблюдаются активная вовлеченность в процесс и заинтересованность. Сотрудники способны делать наблюдения и знают, что за их работой тоже наблюдают. Когда в компании четко определены цели, а сотрудники сосредоточены на основных KPI, им действительно важно, когда эксперимент проваливается или программа «взлетает». Они будут пытаться разобраться в причинах, чтобы улучшить процесс. Поддерживайте этот настрой и не останавливайтесь, если результаты A/B говорят о «провале», — воспринимайте это как процесс обучения, который позволит в будущем выдвинуть более удачную гипотезу.

Управление на основе данных требует гибкости и готовности вносить изменения и на уровне компании: по мере роста и развития компании вы должны быть готовы реорганизовать свои команды специалистов по работе с данными и изменить их место в структуре организации.

Как противостоять HiPPO

Гиппопотамы — одни из наиболее опасных животных в Африке.
Не менее опасны HiPPO в переговорных.

Джонатан Розенберг[1]

Как уже говорилось в предыдущей главе, представители HiPPO не ладят с данными. Они принимают решения на основе собственного опыта, предвзятого мнения и интуиции, *не обращая внимания на имеющиеся в их распоряжении данные*. Это может быть плохо для бизнеса. Один из способов борьбы с этим явлением — сделать процесс принятия решений прозрачным и подотчетным. Если такие сотрудники принимают отличные решения, способствующие росту и развитию бизнеса, что ж, отлично — в конце концов, именно в этом и состоит ваша цель. Однако если качество их решений вызывает сомнения, их стоит попросить изменить подход к работе или указать на дверь. HiPPO оказывают крайне негативное влияние на корпоративную культуру компании, которая стремится действовать на основе данных. Принимаемые ими решения не всегда эффективны, а из-за их статуса в компании эти решения не подвергаются сомнениям. (Если вы помните комментарий, приведенный в предыдущей главе: «В большинстве компаний представлять данные, противоречащие точке зрения или намерениям HiPPO, — прямой путь к увольнению и попаданию в черный список сотрудников».)

[1] URL: https://plus.google.com/+JonathanRosenberg/posts/DaUY9tT8Ev6.

Иными словами, они препятствуют становлению в компании открытой корпоративной культуры, основанной на сотрудничестве, где каждый может предлагать собственные идеи, где сотрудники готовы честно признать: «Я не знаю, но давайте проверим» и где побеждают лучшие, объективные и подтвержденные данными выводы.

Не поймите меня превратно: иногда интуиция и опыт действительно могут играть весьма важную роль. В некоторых случаях у вас просто может не быть данных, особенно если вы действуете в новой области. Иногда данные бывают информативными, но кто-то должен принять окончательное решение, возможно, при наличии неопределенности или неизвестных данных. Говоря о HiPPO, я имею в виду именно тех людей, которые отказываются от использования доступных данных, особенно если раньше они уже принимали неудачные решения и если они ни перед кем не отчитываются, какое решение принимают. Представьте, каково аналитику работать (или бороться?) с таким руководителем. Если данные противоречат управленческим решениям, но руководителя это не волнует, это создает ситуацию противостояния, которая редко заканчивается добром.

Руководство на основе данных

Никто не может сравниться с руководителем, ставящим во главу угла данные и анализ.

Рассел Гласс[1]

В компании, где реализуются принципы управления на основе данных, должна быть сильная вертикаль власти, поддерживающая эти принципы. Руководство должно стимулировать и продвигать соответствующую корпоративную культуру и активно поддерживать все аспекты аналитической цепочки ценности — от сбора данных до принятия решения на их основе и обучения. Руководство должно продвигать методы работы на основе данных.

Подобные принципы руководства позволяют компаниям, по словам Дэвенпорта и его коллег, «конкурировать в аналитике». По результатам недавнего исследования, 58% респондентов из компаний — лидеров в своей области подтвердили, что топ-менеджмент личным примером

[1] Economist Intelligence Unit. The Virtuous Circle of Data: Engaging employees in data and transforming your business (London: Economist Intelligence Unit, 2015). URL: http://live.wavecast.co/virtuouscircleofdata/.

стимулирует развитие в компании корпоративной культуры, ориентированной на данные, по сравнению с 49% в «средних» компаниях или компаниях-аутсайдерах (рис. 10.3). И наоборот, 41% респондентов из компаний-аутсайдеров отметили, что отсутствие поддержки со стороны руководства препятствует более активному использованию данных по сравнению с 23% в компаниях-лидерах.

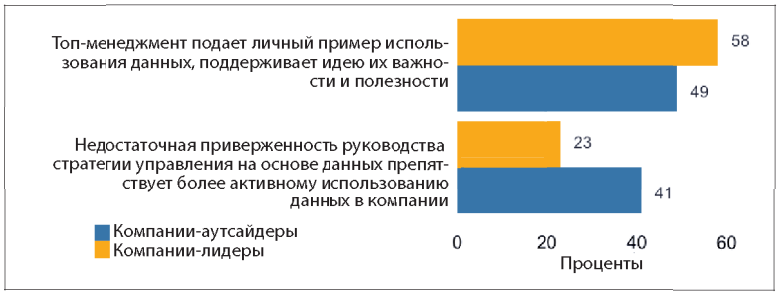


Рис. 10.3. В компаниях, превосходящих конкурентов, выше вероятность сильного руководства

Источник: The Virtuous Circle of Data: Engaging employees in data and transforming your business, аналитическое подразделение журнала Economist (http://live.wavecast.co/virtuouscircleofdata/)

Руководитель, реализующий принципы управления на основе данных, ориентируется на несколько групп.

Во-первых, он должен поддерживать специалистов аналитического отдела. Руководителю следует обеспечить им инструменты и обучение в случае необходимости. Руководитель определяет организационную структуру, меняя ее соответствующим образом по мере роста и развития компании. Кроме того, он должен показать четкую карьерную лестницу и стимулы для специалистов аналитического отдела, чтобы повысить их продуктивность и личную удовлетворенность.

Во-вторых, руководитель должен добиться, чтобы его поддерживали все остальные сотрудники, особенно когда речь идет о коммерческом предприятии. Он должен быть уверен в правильности выбранного им подхода на основе данных. Чтобы заручиться этой поддержкой, руководитель должен демонстрировать результаты, пусть сначала даже небольшие. Благодаря этому у руководителя повысятся шансы на продвижение корпоративной культуры на основе данных, которую будут поддерживать все подразделения компании.

Наконец, руководителя должны поддерживать остальные топ-менеджеры компании. Они отвечают за бюджеты на развитие нужной ИТ-инфраструктуры и обучение, а также играют основную роль в стимулировании корпоративной культуры на основе данных в *своих* подразделениях.

Это поверхностный обзор руководства на основе данных, требующий более глубокого изучения. Так как такое руководство — чрезвычайно важный фактор при продвижении соответствующей корпоративной культуры в компании, этой теме будет посвящена следующая глава.

ГЛАВА 11

Топ-менеджмент компании с управлением на основе данных

Идеальный CDO стимулирует бизнес-возможности.
Джули Стил[1]

Если компания хочет внедрить управление на основе данных, должен быть увлеченный этой темой человек, который будет привлекать внимание к тому, что для этого нужно.
Аноним в сборнике Shaw et al. (2014)[2]

Мы уже рассмотрели достаточно много аспектов. Наше изучение темы проходило в парадигме «снизу вверх» — от данных и влияния необработанных данных до аналитической цепочки ценности. Мы начали с основ — с уровня данных (то есть сбора правильной информации и правильного сбора информации). Затем перешли к структуре аналитического подразделения и поиску профессионалов с необходимыми навыками, которые способны сделать важные аналитические выводы. Далее мы изучили разные типы статистических инструментов и инструментов визуализации, а также подходы с использованием сторителлинга, которые в итоге могут превратить сырые данные в презентации, облегчающие принятие решений. Важный пункт, на который мы обратили отдельное внимание, — что аналитики и их непосредственные руководители могут сделать для

[1] Steele J. Understanding the Chief Data Officer. Sebastopol, CA: O'Reilly, 2015.

[2] Shaw T., Ladley J. and Roe C. Status of the Chief Data Officer: An update on the CDO role in organizations today, Dataversity, November 5, 2014. URL: http://whitepapers.dataversity.net/content42609/.

стимулирования корпоративной культуры на основе данных и для достижения успеха.

Теперь пришло время сменить парадигму и изучить тему с точки зрения вертикали власти. Конечно, в компании с управлением на основе данных корпоративная культура может процветать и приносить плоды, действуя с самых низких уровней организации, но, чтобы полностью реализовать заложенный в ней потенциал, ее должны поддерживать и направлять «сверху», то есть в компании должно осуществляться руководство на основе данных. Этой теме и будет посвящена эта глава.

В компании должен быть топ-менеджер, отвечающий за данные. В течение длительного времени это был CTO (Chief Technology Officer, технический директор) или CIO (Chief Information Officer, директор по информационным технологиям). Однако для них данные не представляли стратегический актив, поскольку эти сотрудники преимущественно концентрировались на информационных *системах* как таковых, то есть на инфраструктуре для поддержания компании на плаву. К счастью, в последнее десятилетие наблюдается определенный сдвиг, так как все больше компаний уже не ассоциируют данные только с расходами и обязательствами, но оценивают их как актив. В свете этого информация и аналитика приобретают все более важное значение и становятся теми аспектами, которыми следует управлять и которые необходимо оптимизировать. В результате появился целый ряд позиций в рамках руководства высшего звена: CDO, CAO и Chief Digital Officer. Если вам кажется, что две позиции с аббревиатурой CDO создают путаницу, должен вам сообщить, что недавно корпорация Johnson & Johnson наняла на работу Chief Design Officer (директора по дизайну)[1], так что появилась третья позиция с этой аббревиатурой.

Эти новые управленческие позиции вызывают много вопросов, а потому я подробнее остановлюсь на двух из них: CDO и директор по аналитике. (Я не буду особо касаться позиции Chief Digital Officer, так как он играет менее важную роль во внедрении в компании управления на основе данных). Для каждой из этих позиций я опишу функционал, историю и персональные качества, необходимые для успеха. Кроме того, я остановлюсь на потенциальном влиянии на компанию и на том, как определить, нужен ли компании один из этих топ-менеджеров, оба или не требуется ни одного.

[1] URL: https://www.wsj.com/articles/SB10001424052702304256404579449290361956838.

Chief Data Officer

CDO — позиция, которая исторически появилась раньше двух остальных. Первым CDO в январе 2002 года была назначена Кэтрин Клей Досс в компании Capital One. С этого момента количество назначаемых CDO начало стремительно расти[1].

Сегодня CDO преимущественно встречаются в следующих областях:

— банковский и финансовый сектор (40% от общего количества);

— государственное управление;

— здравоохранение.

Это распределение уже само по себе может дать некоторое представление о роли CDO[2]. Что объединяет все эти области? Регулирование. Все они подчиняются строгому регулированию на уровне местного самоуправления, штата или на федеральном уровне. Соответствие финансовой отчетности требованиям закона Сарбейнза–Оксли, или исполнение требований Закона США о сохранении медицинского страхования и персонифицированном учете в здравоохранении (HIPAA), или банковские требования в сфере противодействия отмыванию средств — все эти виды деятельности непосредственно связаны с большими данными и представляются сложными, подлежащими непрерывному контролю, а их нарушение сопряжено с серьезным наказанием. Для этих организаций это основной фактор риска.

Однако это еще не всё. Организации, действующие в перечисленных областях, подчинялись требованию сбора и защиты данных задолго до 2002 года. Что же изменилось? Вероятно, пришло осознание, что управлять данными можно иначе, что данные могут быть активом, а не только обязательствами, требующими расходов, и что можно заставить данные работать по-новому. На последнем симпозиуме банковских аналитиков Banking Analytics Symposium в Нью-Орлеане только 15% участников подтвердили, что в их организации есть CDO или

[1] Сегодня в мире насчитывается примерно 200 CDO. По прогнозу исследовательской и консалтинговой компании Gartner, к 2015 году в 25% крупных международных корпораций будет должность директора по большим данным. Шоу и др. предполагают, что число CDO будет удваиваться примерно каждые 15–18 месяцев в течение следующих пяти лет. URL: http://www.gartner.com/newsroom/id/2659215.

[2] Хотя постепенно эта управленческая позиция начинает появляться в компаниях, действующих в таких областях, как информационные услуги, страхование, электронная коммерция (Shaw et al., 2014), а также медиа и производство (цит. по Д. Велланте). URL: https://www.youtube.com/watch?v=_LeVQ8yw4t4.

аналогичная должность. Чарльз Томас[1], вновь назначенный CDO банковской компании Wells Fargo, выступая на этом симпозиуме, отметил: «Вскоре этот тренд станет более заметным [в банковском секторе], так как фактически мы сидим на тоннах данных и не используем их должным образом»[2].

Таким образом, основная обязанность CDO (или она должна быть таковой) заключается в стратегическом использовании данных. Марио Фариа, один из первых CDO, как-то сказал мне: «Лучшие из CDO занимаются не только контролем и управлением. Они стимулируют бизнес-возможности и через свою команду реализуют новые способы использования данных для потребностей бизнеса». Учитывая сказанное, давайте подробнее остановимся на роли CDO.

РОЛЬ CDO

IBM определяет CDO как «руководителя, разрабатывающего и реализующего стратегии работы с данными и стратегии аналитической работы для стимулирования бизнес-возможностей»[3]. Таким образом, зона ответственности CDO получается довольно обширной и охватывает как технические, так и нетехнические аспекты. Обратите внимание, что спектр обязанностей, который мы будем обсуждать, идеализированный и весьма условный. Вряд ли вам удастся найти двух CDO с одинаковым набором обязанностей, так как все зависит от конкретной ситуации в компании: бюджета, персонала, формы отчетности (обо всем этом мы поговорим далее).

Одна из возможных функций CDO заключается в наблюдении за информационными технологиями по работе с данными или в управлении ими. CDO определяет ви́дение, стратегию, процессы и методы, посредством которых в компании осуществляются сбор, хранение и управление данными, а также контроль их качества. Это подразумевает управление персоналом, например специалистами по работе с данными. Как отмечалось в главе 2, это основополагающий компонент,

---

[1] URL:https://www.information-management.com/news/chief-data-officers-battle-complexity-complacency-wells-thomas.

[2] Crosman P. Chief Data Officers Battle Complexity, Complacency: Wells' Thomas, Information Management, October 30, 2014. URL: https://www.information-management.com/news/chief-data-officers-battle-complexity-complacency-wells-thomas.

[3] IBM Software, Insights for the New Chief Data Officer, IBM Corp., June 2014. URL: https://www.information-management.com/news/chief-data-officers-battle-complexity-complacency-wells-thomas. См. Также: The Role of Chief Data Officer in the 21st Century. URL: https://www.cutter.com/article/role-chief-data-officer-21st-century-400806.

и его отсутствие может привести к некачественным исходным данным и сомнительному итоговому результату.

В обязанности CDO часто входит контроль над определением стандартов и политики деятельности. Это может быть довольно широкий круг — от качества данных и обмена информацией до определения уровней доступа к данным. Кроме того, CDO отвечает за разработку и поддержание словарей данных и обеспечение доступа к ним во всей компании. Это основной компонент, позволяющий избежать путаницы с принятой в компании терминологией и убедиться, что все сотрудники говорят на одном языке. Важность этого шага трудно переоценить. В компании Warby Parker моя команда тесно работала с руководством для определения словаря данных, его документирования и реализации этих четких бизнес-правил в основном инструменте бизнес-аналитики[1]. Возможно, это самое важное из того, что удалось нам сделать вплоть до настоящего момента, так как это позволило устранить путаницу, позволило проводить рациональные сравнения показателей и обеспечило создание надежного единого источника данных внутри компании. В успешной компании с управлением на основе данных бывает множество проектов, связанных с данными, в реализации которых принимают участие как команда специалистов под руководством CDO (если он есть), так и команды других подразделений. Таким образом, роль CDO должна заключаться в осуществлении поддержки этой деятельности путем управления, координирования и следования общей стратегии. Кроме того, CDO должен измерять и контролировать эффективность этих проектов, стимулируя работу для получения максимального эффекта и рентабельности затраченных аналитических усилий.

CDO может осуществлять руководство аналитическим подразделением, контролируя команды аналитиков и/или специалистов по работе с данными. Но если нет, он в любом случае непосредственно взаимодействует с участниками и руководителями этих команд. Все эти ресурсы стоят денег, так что CDO может иметь в своем распоряжении бюджет, который расходуется, например, на покупку программ по повышению качества данных, привлечение высокопрофессиональных аналитиков, обучение, покупку данных для дополнения доступных внутренних данных и так далее.

Основная функция CDO — определение и изучение новых бизнес-возможностей. Это включает как возможность генерировать новые

[1] Anderson C. Creating a Data-Driven Organization: Two Years On, April 6, 2015. URL: http://www.p-value.info/2015/04/creating-data-driven-organization-two_6.html.

источники прибыли, так и развитие бизнеса в новых направлениях. В зависимости от положения CDO в компании, бюджета и ресурсов, которыми он располагает, это может означать как изучение самих идей, так и обеспечение возможностей для других команд изучить данные, результаты визуализации данных и другие продукты на основе данных.

Какой тип возможностей окажется рациональным, зависит преимущественно от сферы деятельности компании и ее бизнес-модели. Например, в области государственного управления, где деятельность CDO сконцентрирована на обеспечении прозрачности и публичной подотчетности, это может означать определение и обеспечение доступности тех наборов данных, которые имеют ценность для других городов, штатов или граждан в целом. Желательно, чтобы эти данные были в формате, доступном для обработки с помощью программных кодов[1]. Уровень успеха определяется тем, что другие применяют ваши данные и извлекают из них пользу в качестве всеобщего блага. Для многих компаний успех может означать возникновение инноваций в результате использования данных, которыми они располагают на данный момент. Выступая в Нью-Йорке на конференции Strata+Hadoop World 2014, заместитель министра торговли по вопросам экономической деятельности Марк Домс рассказал, что доля ответивших на опросы в ходе переписей населения США составляет 88%. Чтобы повысить этот процент, нужно ходить по домам, что очень дорого. Чтобы максимально повысить эффективность этих визитов, они дополнили данные переписи данными из программы по социальному страхованию, чтобы оценить, кто должен быть дома и в какое время.

Другие компании занимаются бизнесом по сравнению данных, их дополнению и продаже. Для этих компаний успех определяется возможностью найти новые источники данных, по-новому дополнить данные и предоставить специалистам по продажам информацию о новых товарах, которые могут иметь ценность для их клиентов.

Все больше компаний из сферы маркетинговых услуг и работы с данными начинают вводить должность CDO. Выступая на форуме Chief Data Officer Executive Forum в Нью-Йорке, Мэттью Грэйвз, занимающий эту позицию в компании InfoGroup, обозначил суть роли CDO — евангелизм, то есть продвижение и популяризация[2], даже если речь идет

[1] Компьютеры могут взаимодействовать и обмениваться данными посредством интерфейсов программирования приложений (APIs).

[2] ИТ-евангелист (ИТ-пропагандист) — специалист, профессионально занимающийся пропагандой в сфере информационных технологий. Как правило, это человек, который

о компании, занимающейся продажей данных. Нужно образовывать сотрудников компании, внутренних специалистов по продаже данных, клиентов, объяснять им суть улучшений, которые происходят в области работы с данными, и давать информацию о новых данных. Клиенты не привыкли использовать данные, и в этом главная причина, по которой компании, торгующие данными, стремятся ввести должность CDO.

Чтобы внедрить управление на основе данных, компания должна начать относиться к данным как к стратегическому активу. А чтобы этого добиться, необходимо стимулировать всех сотрудников: им нужны конкретные примеры, показывающие, как данные влияют на их деятельность и повышают ее эффективность. Поэтому CDO должен обладать хорошими навыками коммуникации и способностью разговаривать с ИТ-специалистами на одном с языке, чтобы мотивировать их и вдохновлять.

CDO должен менять корпоративную культуру, оказывать влияние на других (как на топ-менеджеров, так и на простых сотрудников), чтобы изменить их отношение к использованию данных. Ему необходимо способствовать созданию в компании открытой корпоративной культуры, основанной на обмене данными, а также демократизировать данные, делая доступными их источники, что включает ликвидацию обособленных закрытых хранилищ данных[1]. Иными словами, он должен повысить уровень доступности данных и усовершенствовать умение обращаться с ними в компании в целом. Это масштабная и очень серьезная задача.

СЕКРЕТЫ УСПЕХА

Директор по большим данным — это, главным образом, евангелист и агент изменений. Как однажды заметил Питер Айкен, соавтор книги The Case for the Chief Data Officer: «Никто не привлекает CDO, если дела в компании идут хорошо». Если так, то что необходимо для успеха? Конечно, требуется совокупность технических навыков и социальных компетенций. Например, когда я спросил Марио Фариа, какие навыки необходимы CDO, он ответил: «Мы должны совмещать технические навыки (опыт работы с данными, техническую и статистическую

аккумулирует вокруг себя некоторую массу людей с целью создания целевой аудитории для продвижения продукта на рынке и утверждения его как технологического стандарта с возможностью возникновения сетевого эффекта. *Прим. перев.*

[1] Конечно, в этом правиле есть исключения. Джон Минкофф — CDO бюро по обеспечению исполнения Федеральной комиссии по связи США. Его команда работает в основном с данными обвинительных решений, и ни у одного другого бюро ФКС нет доступа к их данным, что вполне объяснимо.

грамотность, профессиональные знания, деловую хватку) и социальные компетенции (навыки коммуникации, управления, уважение разнообразия, стремление изменить существующее положение дел).

Питер Айкен[1] провел опрос[2] среди руководителей, применяющих принципы управления на основе данных, и выяснилось, что тремя главными качествами CDO они считают:

— сбалансированную совокупность технических навыков, знаний в области ведения бизнеса и социальных компетенций;
— отличные навыки коммуникации и выстраивания взаимоотношений;
— стратегическую подкованность (с позиции политики компании).

Очевидно, что это выходит за рамки только технической роли.

Подотчетность CDO

Итак, кому иерархически подчиняется CDO? В идеале он отчитывается в своих действиях перед CEO и занимает равное положение с другими топ-менеджерами: CTO, CIO, CFO (Chief Financial Officer, финансовым директором), COO (Chief Operating Officer, операционным директором), CISO (Chief Information Security Office, директором по информационной безопасности) и так далее. Однако на практике в 80% случаев CDO подчиняется непосредственно CTO (цит. по Айкену, на основе его опроса 2013 года[3]).

Что плохого в подчинении техническому директору? Айкен (с. 52) утверждает:

> CDO неспособен обеспечить использование данных, если иерархически он находится в подчинении у технического директора. Более того, если до первых лиц компании

[1] Aiken P. The Precarious State of the CDO: Insights into a burgeoning role, Data Blueprint, July 16, 2013.

[2] URL: http://datablueprint.com/publications/2013-The-Precarious-State-of-the-CDO.pdf.

[3] Шоу и др. (2014) утверждают, что «CDO в большинстве случаев подчиняется генеральному или операционному директору или другому первому лицу компании. Очень немногие CDO подчиняются директору по информационным технологиям, а скорее занимают равную с ним позицию». Возможно, ситуация значительно изменилась за один год. Тем не менее следует учитывать, что размер выборки Шоу существенно меньше, а значит, здесь может иметь место эффект размера выборки, «ошибка выжившего» (опрашиваемые Шоу специалисты были более успешными и имели бóльшую степень поддержки) или другие факторы.

> результаты его работы доносит человек, не обладающий навыками работы с данными, то улучшить процесс принятия решений оказывается практически невозможно.

Автор полагает, что в большинстве случаев технические директора не обладают нужными навыками по управлению данными, закрыты и придерживаются иного взгляда на управление проектами. По его словам, «работа с данными происходит в другом ритме, нежели работа с программным оборудованием, и ее нельзя рассматривать как проект. Управление данными должно осуществляться на программном уровне. В противном случае у данных должны быть начало и конец, а с ними так не получается».

Иными словами, данные могут поддерживать несколько проектов одновременно, и, поскольку они составляют основу проектов, то часто выходят далеко за их границы. Таким образом, CDO смогут принести компании больше пользы, если будут подчиняться людям, отвечающим за коммерческую составляющую, а не за техническую.

Мандат на влияние

Самое поразительное открытие, следовавшее из опроса Айкена, состоит в том, что «почти половина CDO не располагают бюджетом, у половины из них нет сотрудников в подчинении, более 70% не имеют нужной поддержки на организационном уровне».

С такими скудными ресурсами CDO фактически остается только роль евангелиста и лидера группы поддержки. К сожалению, на голом энтузиазме долго не продержишься. В конце концов, от него ждут результатов, добиться которых фактически можно только при наличии команды и бюджета. Признавая это, компания Gartner[1] предполагает[2], что «людям, занимающим пока еще новую должность CDO, придется столкнуться с серьезными вызовами и конфликтующими приоритетами, так как для этой роли в компании пока еще не определена профессиональная структура и нет самых эффективных методов работы».

Чтобы успешно справляться со своим функционалом, даже если у вас нет команды или бюджета, вы должны обладать полномочиями принимать решения. Марк Хэдд, первый CDO в Филадельфии, успешно

---

[1] URL: https://www.gartner.com/doc/2648615/cio-advisory-chief-data-officer.

[2] Logan D. and Raskino M. CIO Advisory: The Chief Data Officer Trend Gains Momentum, January 13, 2014. URL: https://www.gartner.com/doc/2648615/cio-advisory-chief-data-officer.

выпустил ряд массивов данных, но столкнулся с непреодолимым препятствием в виде API, связанного с налогом на имущество, который выплачивается в городской бюджет[1]. Он встретил серьезное сопротивление со стороны представителя налогового управления. Вот что рассказывает Марк:

> Филадельфия стояла на перекрестке и была готова сделать следующий шаг в направлении эволюции данных. Мы были готовы начать обмениваться данными между департаментами (да что там, даже между органами управления) и находить новые, более эффективные способы ведения деятельности. Я приложил все мыслимые усилия, чтобы этот перекресток был пройден в верном направлении. Мне это не удалось, и теперь очевидно, что у меня никогда бы этого не получилось. Механизм самостоятельной сертификации через сайт — это ответ XX века на проблему неплательщиков налогов. Но XXI век предлагает новое решение этой проблемы — открытый интерфейс программирования приложений (API). Это стало моим сильнейшим разочарованием за время работы в этой должности: мы постоянно применяли решения прошлого века для тех проблем, где требовался новый подход.

Менять отношение и общую культуру очень непросто.

Когда Джон Боттега пришел на работу в Bank of America, у него уже был блестящий послужной список: CDO в обоих Citi (2006–2009) и в Federal Reserve Bank of New York (2009–2011). Он рассказывает: «Когда большинство компаний только вводили должность CDO (хотя называться она могла по-разному), это казалось отдельным и обособленным направлением в бизнесе. Сегодня это скорее горизонтальная функция, которая распространяется на всю компанию». С учетом сказанного, у Боттеги не было организационной структуры, с которой он мог бы начать работать и определить зону ответственности, и у него практически отсутствовала поддержка. Более того, он попал в очень сложную ситуацию. Bank of America — это огромная организация (более 200 тыс. сотрудников), а отдельные направления бизнеса — сами по себе отдельный бизнес: управление активами, депозиты, ипотечное кредитование, кредитные карты и так далее. «Если вы стремитесь

[1] Reyes J. Why Philadelphia's first Chief Data Officer quit, Technical.ly Philly, June 19, 2014. URL: https://technical.ly/philly/2014/06/19/why-philadelphia-chief-data-officer-quit/.

объединить сотрудников вокруг корпоративной цели или задачи, сделать это нереально сложно», — признаётся Питер Прэсланд-Брин, на тот момент старший вице-президент, главный архитектор подразделения Bank of America по жилой недвижимости. «Представьте, что вы пришли в Bank of America на должность CDO. Это позиция корпоративного уровня, и предполагается, что вы будете влиять на все направления бизнеса, которые и без того успешны и получают свое вознаграждение независимо от действий CDO». Должность, которую занимал Боттега, упразднили всего через два года[1].

Конечно, можно привести и противоположные примеры. Некоторые CDO располагают бюджетом, ресурсами и поддержкой, необходимыми для достижения успеха. В распоряжении Чарльза Томаса из Wells Fargo, по его словам, «скромная команда» из 600 человек и бюджет в 10 млн долл. У Кайла Эванса, CDO компании RP Data, в подчинении 200 человек. У Мишелин Кейси, CDO совета управляющих Федеральной резервной системы США (и CDO штата Колорадо в период с 2009 по 2011 год) команда из 25 человек, а операционный бюджет в 2014 году составил примерно 10 млн долл. «Если почитать стратегический документ Federal Reserve Board (Федеральная резервная система США)[2], это просто идеальная работа для CDO», — говорит Льюис Брум. CDO подчиняется COO, который, в свою очередь, подчиняется председателю совета. Более того, инициатором введения этой должности был именно председатель совета. Таким образом, требуется поддержка CEO и совета директоров и понимание того, что управление данными — один из стратегических приоритетов для компании, для реализации которого нужен руководитель уровня топ-менеджера, а также бюджет и всесторонняя поддержка.

Еще одна стратегия достижения успеха заключается в том, чтобы найти единомышленников. Для Грега Элина, первого назначенного CDO Federal Communications Commission (FCC, Федеральная комиссия по связи США), таким соратником стал Майкл Брин. «Майкл, который был первым в истории FCC GIO (Geographic Information Officer, директор по географической информации), как и я, верил в эффективность

[1] По словам Питера, команда по работе с данными начала наращивать обороты, когда Bank of America сконцентрировался на коммерческой ценности, особенно на углублении взаимоотношений с клиентами. С тем посылом, который шел от главы банка Брайна Мойнихэна, у сотрудников были причины и стимулы стремиться к работе с качественными данными, обмену информацией и управлению на основе данных.

[2] Federal Reserve Board. Strategic Framework 2012–2015, 2013. URL: https://www.federalreserve.gov/publications/gpra/2013-strategic-themes.htm#subsection-153-AC33F9CB.

RESTful APIs, — рассказывает Грег. — Хочется верить, что именно наличие этих двух ключевых позиций, CDO и GIO, продвигавших новый, особенный подход к работе с данными, повлияло на появление интерфейсов программирования приложений в проекте такого уровня [создание National Broadband Map — Национальной карты широкополосного доступа] в агентстве, где подобные инструменты никогда не использовались».

ПЕРВЫЕ 90 ДНЕЙ

Я попросил Марио Фариа рассказать о его стратегии поведения в первые три месяца после вступления в должность:

> Первые 90 дней очень важны, особенно если вы пришли в новую компанию. Первый месяц стоит потратить на то, чтобы как можно больше общаться с сотрудниками — от топ-менеджмента до стажеров. Вы должны понять, что происходит в компании, и начать выстраивать свои стратегические связи.
>
> В течение второго месяца определитесь со своими краткосрочными, среднесрочными и долгосрочными планами. Помимо этого, в это время вам следует сформулировать миссию и ви́дение для компании. На основе этого вы сможете понять, как вашей команде действовать дальше.
>
> На третий месяц, после того как ваш план готов и получил одобрение, приступайте к реальным действиям. Самое время начать добиваться пусть небольших, но положительных результатов. Вы должны продемонстрировать прогресс своей команде, чтобы их мотивировать, и всей остальной компании, чтобы доказать, что принять вас на работу было верным решением.

Грег Элин:

> Для меня самыми важными были евангелизм и поиск очевидных возможностей повысить эффективность сбора данных, их использования, управления данными и распределения, чтобы стимулировать изменения в отношении сотрудников к данным и работе с ними. На момент, когда я стал CDO, в агентстве уже реализовывались важные проекты на основе данных. Уже был объявлен Национальный план развития широкополосного доступа, и FCC проводила тестирование широкополосного доступа. Осуществлялась

разработка национальной карты широкополосного доступа. Руководитель FCC требовал обзор всех массивов данных с учетом и обоснованием затрат и потребностей, и мы готовили список массивов данных от трех основных бюро для вынесения его на публичное обсуждение: что стоит сохранить, что изменить, а от чего избавиться. Так что мне пришлось выполнять множество срочных задач и по ходу дела оценивать, как в агентстве обстоит дело с работой с данными.

Моя личная стратегия чем-то напоминает стратегию Марио: я беседовал со многими сотрудниками из разных подразделений, чтобы понять текущую ситуацию, систематизировать разные источники данных и оценить их относительную важность. В каждом подразделении я задавал два основных вопроса. Первый: в чем вам требуется помощь с текущими данными и процессами? Это помогало мне определить болевые точки и понять, что нужно сделать в первую очередь для достижения быстрых результатов. Второй вопрос был направлен на перспективу: что вы не в состоянии делать сейчас, с чем мы вам можем помочь? Это помогало определить новые источники данных или недостающую функциональность, на основе чего можно было строить долгосрочные планы работы.

БУДУЩЕЕ ДОЛЖНОСТИ CDO

Если CDO — это преимущественно агент изменений, занимающийся продвижением культуры работы с данными, что станется с этой должностью, когда цель будет достигнута? Будет ли по-прежнему необходимость в этом специалисте? Айкен (2013, с. 65) высказывается в пользу того, что эта должность может быть вре́менной, и проводит параллель с должностью директора по электрификации. Эту должность можно было встретить в организациях примерно в 1880-е годы, в эпоху перехода от использования пара к применению самой современной технологии — электричества. Конечно, сегодня электричество стало нашей повседневной реальностью, а сама должность постепенно исчезла в 1940-е годы. Возможно ли, что нечто похожее произойдет с данными, а также с ролью CDO?

Грег Элин в основном согласен:

> Думаю, роль CDO, как она понимается сейчас, то есть определенного топ-менеджера, отвечающего за преобразование данных в актив для повседневного использования, должна исчезнуть в ближайшие пару десятков лет, поскольку использование данных и проведение анализа

> станут неотъемлемой частью ведения бизнеса. Компании выигрывают больше всего, когда и данные, и ИТ-направление в целом развиваются как часть бизнес-процессов. Иными словами, сегодня CDO должен сосредоточиться на развитии возможностей данных и даже их самодостаточности в рамках всей компании.

К сожалению, как и со многими другими должностями, которые были созданы в компаниях для решения конкретных проблем, должность CDO может закрепиться в структуре организации и после того, как все соответствующие проблемы будут решены. Эта должность имеет важное значение сейчас, поскольку компаниям оказалось легче продвигать и внедрять изменения, когда за них отвечает конкретный назначенный человек.

Ричард Стэнтон, CDO компании Penton Media, более категоричен:

> Нет никаких сомнений в том, что роль CDO станет еще более важной. Не знаю, как именно она будет называться, но ее функционал — тот спектр вопросов, за который сейчас несет ответственность человек на этой позиции, — будет присутствовать в каждой организации. Я абсолютно в этом уверен[1].

Кортни Амберкромби (Emerging Roles Leader в корпорации IBM) в разговоре с Дейвом Велланте (соруководителем SiliconANGLE) отметила:

> Больше чем уверена, что позиция CDO никуда не денется. Более того, она заставит некоторые другие должности видоизмениться, поскольку сегодня данные приобретают очень большое значение для конкурентного преимущества компаний. Это на самом деле новый способ внедрять инновации, лучше узнавать свои сегменты покупателей. Я не вижу предпосылок для упразднения этой должности, ее важность будет только расти[2].

Дейв поддержал эту точку зрения:

> Согласен. Особенно в сферах деятельности с жестким регулированием. Это станет нормой.

---

[1] URL: http://whitepapers.dataversity.net/content42609.

[2] URL: https://www.youtube.com/watch?v=_LeVQ8yw4t4.

При этом Льюис Брум более осторожен:

> Я не уверен, что люди настолько хорошо изучили данные, чтобы понять, нужна эта должность или нет.

После этого краткого обзора роли CDO давайте сравним ее с ролью CAO.

Chief Analytics Officer

Функционал CDO и CAO в значительной степени перекликается. Но если первый фокусируется в большей степени на бэкэнде (то есть на управлении данными), то второй сосредоточен на стратегическом использовании данных, то есть, как следует из названия этой должности, на их анализе. Если в подчинении CDO *могут* быть специалисты по аналитике, то у CAO они *обязательно* должны быть.

По словам Фрэнка Бина, CEO компании Looker, «данные имеют стратегический характер, только если их проанализировали, поняли и на их основе начали предпринимать действия в рамках всей компании, так что ценность этих данных была полностью реализована»[1].

Итак, мы добрались до основной мысли книги. Роль CAO — повысить эффективность методов работы и корпоративной культуры на основе данных и принести ощутимую пользу компании. Билл Фрэнкс, CAO компании Teradata, утверждает[2]:

> По мере усложнения технологий компании все отчетливее начинают понимать силу того, что делает аналитика. Появление должности директора по аналитике — естественное расширение этого процесса, потому что чем больше аналитика укореняется на всех уровнях организации, тем выше потребность в топ-менеджере, который будет отвечать за этот стратегический аспект.

CAO должен обладать способностью разглядеть потенциал в имеющихся данных, понять, как они соотносятся, и объединить все разрозненные источники данных из разных подразделений оптимальным образом. Кроме того, он должен контролировать деятельность аналитической

[1] Bien F. It's Time To Welcome The Chief Analytics Officer To The C-Suite, Fast Company, July 28, 2014. URL: https://www.fastcompany.com/3033590/the-future-of-work/its-time-to-welcome-the-chief-analytics-officer-to-the-c-suite.

[2] O'Regan R. Chief analytics officer: The ultimate big data job? Computerworld, October 3, 2014. URL: http://cw.com.hk/feature/chief-analytics-officer-ultimate-big-data-job.

структуры компании, обеспечивать обучение и повышение квалификации и при необходимости проводить реорганизацию. Как правило, это означает централизацию[1] в рамках модели центра компетенций, а также интегрированной или гибридной модели. То есть, вероятно, должность директора по аналитике вводится тогда, когда специалисты по аналитике уже работают в разных частях компании, например в составе отдельных бизнес-единиц. Как отмечалось ранее (глава 4), эта модель ведет к недостатку стандартов, избытку усилий и неопределенному карьерному пути для специалистов по аналитике. Благодаря централизации всей аналитической деятельности под руководством одного лидера компания получает экономию от масштаба, стандартизацию процессов, кроме того, повышается качество работы и степень удовлетворенности работой у команды аналитиков.

Как сказано в одном из отчетов[2], «директор по аналитике не просто руководитель, а человек, который больше всех остальных требует генерирования ценности из данных. В качестве топ-менеджера он должен обеспечить, чтобы полученные аналитические выводы ложились в основу постоянных действий. Кроме того, он должен лучше технических руководителей понимать, куда и как направить компанию в бурных водах больших данных и большой аналитики». Компания Sandhill Group подготовила доклад под названием Mindset over data set: a big data prescription for setting the market pace[3], в котором выделены следующие качества директора по аналитике (см. с. 275).

Как видите, часть тех качеств, которые приписывают директору по аналитике, можно отнести к качествам евангелиста *больших данных*. В моем исследовании практически каждое описание роли CAO содержало этот компонент. Конечно, это отражает текущую ситуацию и шумиху по поводу темы больших данных[4].

---

[1] Rajaram D. Does Your Company Need A Chief Analytics Officer? Forbes, August 8, 2013. URL: https://www.forbes.com/sites/ciocentral/2013/08/08/does-your-company-need-a-chief-analytics-officer/.

[2] Akmeemana C., Stubbs E., Schutz L. and Kestle J. Do You Need a Chief Analytics Officer? Ontario: Huntel Global, 2013. URL: http://www.huntelglobal.com/wp-content/uploads/HG_Whitepaper_CAO-LowRes.pdf.

[3] Netke S. and Rangaswami M. R. Selecting a Chief Analytics Officer — You Are What You Analyze, SandHill Group, March 3, 2014. URL: http://sandhill.com/article/selecting-a-chief-analytics-officer-you-are-what-you-analyze/. У меня не было лишних 1995 долл., чтобы прочитать полную версию доклада.

[4] Согласно графику развития новых технологий от компании Gartner «Hype cycle for emerging technologies» в 2014 году, большие данные практически в шаге от того, чтобы покинуть «Пик чрезмерных ожиданий» и опуститься в точку «Избавление от иллюзий». URL: http://www.gartner.com/newsroom/id/2819918.

Характеристика	Качества
Аналитик	Обладает отличными навыками анализа данных и бизнеса; формулирует и оценивает рациональные показатели эффективности деятельности; обладает обширным опытом аналитической работы, что вызывает уважение и у команды, и у руководящего звена.
Евангелист	Помогает другим осознать ценность больших данных и многочисленные способы их применения; приводит примеры, вдохновляющие остальных на использование больших данных.
Исследователь	Обладает внутренним любопытством, которое заставляет его задавать интересные бизнес-вопросы и находить инновационные и эффективные решения.
Руководитель	Формирует и поддерживает высокоэффективную команду, умеющую работать с большими данными; заручается поддержкой и активным участием других организаций; эффективно управляет командами как непосредственно, так и в матричных структурах; разрабатывает реалистичные планы, проводит оценку необходимых ресурсов и расходов.
Прагматик	Стремится выявить возможные ошибки на раннем этапе; устанавливает реалистичные ожидания по поводу сроков и результатов; добивается согласия между группами с конкурирующими интересами и приоритетами; поддерживает рациональную расстановку приоритетов в ходе всего проекта.
Технический специалист	Разбирается в применяющихся технологиях и может выступать как равноправный партнер CTO, CIO и CISO.

Я придерживаюсь мнения, что ценность данных не измеряется их масштабом и что сегодня кажется большим, завтра может стать маленьким. Технологии и терминология изменятся, но CAO должен побудить топ-менеджмент и остальных сотрудников компании осознать силу широких, глубоких, дополненных, качественных данных, *обладающих контекстом*, — таких, как в случае с покупкой садовой мебели Белиндой Смит в примере из главы 3. Данные, имеющие контекст, — настоящая

основа прогнозных моделей и рекомендательных сервисов с высокой эффективностью, а также всех более высоких уровней аналитики (глава 1). Задача директора по аналитике — способствовать тому, чтобы это было реализовано на практике.

Как и в случае с CDO, CAO должен заручиться поддержкой первых лиц компании. В настоящее время эта позиция редко относится к высшему руководящему звену, скорее, CAO подчиняется кому-то из топ-менеджеров, отвечающих за коммерческий аспект. По словам Билла Фрэнкса[1], «CAO должен сохранять нейтралитет — этакая Швейцария управленческого звена. Он должен подчиняться топ-менеджеру, отвечающему за все бизнес-единицы, у которого есть потребность в аналитических данных, например Chief Strategy Officer (директор по стратегическому развитию), CFO или COO». Иногда легче бывает сказать, кому *не должен* подчиняться CAO. Например, аналитика маркетинговых данных очень важна для многих компаний. Однако если CAO будет находиться в подчинении у Chief Marketing Officer (директор по маркетингу), остальные бизнес-единицы, например занимающиеся разработкой продукта или обслуживанием клиентов, могут счесть, что их отодвинули на второй план.

Должность CAO появилась позже, чем должность CDO. Согласно одному из отчетов[2], 4 ноября 2013 года 477 пользователей социальной сети LinkedIn указали CDO как название своей текущей должности, в то время как 298 пользователей отметили, что их текущая должность — CAO. (Предположительно, это глобальные цифры, но они все равно кажутся высокими. В декабре 2014 года я обнаружил 357 CDO и 248 CAO при осуществлении глобального поиска с фильтром «только текущая позиция». В США результаты были 181 и 171 соответственно, что совпадало с информацией Gartner.) Больше всего директоров по аналитике было в таких областях, как здравоохранение, медиа и финансовые услуги.

Суть работы CAO, как и CDO, — в том, чтобы стимулировать изменение корпоративной культуры. Добиться этого чрезвычайно сложно, и нередко приходится преодолевать серьезное сопротивление. Необходимо заручиться поддержкой всех бизнес-единиц. Стоит ли удивляться, что на этой «войне» не обходится без потерь. В одной

---

[1] Franks B. Do You Know Who Owns Analytics at Your Company? Harvard Business Review, September 23, 2014. URL: https://hbr.org/2014/09/do-you-know-who-owns-analytics-at-your-company.

[2] Akmeemana C., Stubbs E., Schutz L. and Kestle J. Do You Need a Chief Analytics Officer? Ontario: Huntel Global, 2013. URL: http://www.huntelglobal.com/wp-content/uploads/HG_Whitepaper_CAO-LowRes.pdf.

из телекоммуникационных компаний руководители бизнес-подразделений очень медленно проходили обучение и внедряли модели удержания клиентов и модели ценообразования, разработанные после прихода в компанию нового руководителя аналитического направления. По словам консультантов McKinsey, они «не видели потенциала, который, откровенно говоря, не входил в "их" стратегические приоритеты. По нашему опыту, в большинстве компаний 90% средств вкладывается в разработку моделей и только 10% — в то, чтобы эти модели действительно использовались при работе с клиентами, хотя на практике именно во второй вид деятельности следует инвестировать до половины средств на аналитическую работу». Директор по аналитике должен инвестировать время, средства и усилия в «последнюю милю», побуждая сотрудников, которые непосредственно взаимодействуют с клиентами и при этом пользуются инструментами бизнес-аналитики, а также руководителей этих сотрудников осознать ценность этих инструментов. Он должен обучать сотрудников максимально извлекать эту ценность. Иными словами, это самое слабое звено аналитической цепочки ценности, и его следует укреплять.

Один из подходов, который оправдал себя, по крайней мере, с одной из компаний, производящих товары широкого потребления, состоял в привлечении CEO как силы воздействия. По его указанию руководитель направления по работе с данными и руководитель бизнес-подразделения, который имел весьма слабое представление о больших данных, должны были совместно разработать план для максимальной реализации потенциала аналитических данных. «В результате этого сотрудничества, объединившего эксперта в области данных и опытного руководителя в сфере работы с клиентами, аналитические цели, обозначенные в плане работы, были сконцентрированы на действительно важных и актуальных бизнес-решениях. Более того, когда о результатах этого сотрудничества стало известно остальным топ-менеджерам, эта модель стала активно применяться для планирования деятельности других подразделений». Иными словами, то, что руководителя аналитического направления и представителя конечных бизнес-пользователей стимулировали на столь тесное сотрудничество и, что важно, наделили совместной ответственностью за успех данного предприятия, привело к тому, что их усилия были очень узко сконцентрированы на рентабельности от вложений и оказании реального влияния.

Если ситуация с должностью CDO неопределенна, то, я уверен, должность CAO ожидает светлое будущее. Даже если вскоре компании

начнут получать данные на всех уровнях, им все равно не обойтись без команды специалистов по работе с этими данными, которые будут задавать правильные вопросы, фильтровать информацию и интерпретировать аналитические выводы[1], а также взаимодействовать с теми, кто принимает решения. У этой команды должен быть руководитель, например директор по аналитике.

CHIEF DIGITAL OFFICER

Chief Digital Officer — еще одна новая должность в обойме руководителей высшего звена. Впервые эта должность появилась на телеканале MTV в 2005 году. Основная функция Chief Digital Officer состоит в контроле над реализацией стратегии цифрового развития. Он в меньшей степени ориентирован на внедрение в компании корпоративной культуры на основе данных, и я пишу здесь о нем по двум причинам. Во-первых, функции этого руководителя часто путают с функциями CDO. Во-вторых, одна из задач Chief Digital Officer заключается в стимулировании таких изменений в компании, чтобы она успешно отвечала новым требованиям современной цифровой эпохи. Это серьезно отражается на доступных источниках данных, особенно на типах, характере и разнообразии взаимодействия с пользователями и клиентами. Эти новые потоки данных, часто связанные с местоположением через мобильные устройства, обеспечивают аналитикам богатый дополнительный контекст, а также новые точки контакта и источники взаимодействия, через которые можно предложить продукты на основе данных, например рекомендации, одобрение кредита в режиме реального времени и другие сервисы.

Количество Chief Digital Officer ежегодно удваивалось в период 2005–2013 годов (см. Chief Digital Officer Talent Map[2]) и сегодня достигает более тысячи человек. Фактически их количество превышает совокупное количество CDO и CAO. По мере того как количество мобильных устройств и объем их применения растут, активно развивается интернет вещей[3], характер нашего взаимодействия в цифровом мире быстро меняется. Задача Chief Digital Officer — понять и отследить эти изменения, определить новые сервисы и цифровые предло-

[1] Могу предположить, что бóльшая часть этих выводов будет сгенерирована автоматически, посредством алгоритмов машинного обучения, усиленных еще более сложными технологиями, например такими, как методы глубокого обучения следующего поколения.

[2] URL: http://cdoclub.com/publications/.

[3] Интернет вещей (от *англ.* Internet of Things, IoT) — концепция вычислительной сети физических предметов («вещей»), оснащенных встроенными технологиями для взаимодействия друг с другом или с внешней средой, исключающая из части действий и операций необходимость участия человека. *Прим. перев.*

жения, которые может обеспечить компания, а также выявить новые способы привлечения клиентов. Он понимает, как и когда перевести маркетинговые расходы из аналоговой плоскости в цифровую (с четкой целевой аудиторией) и эффективно использовать социальные сети. Важно, что он помогает связать все эти взаимодействия через разные мобильные устройства в единый, целостный опыт, как со стороны пользователя, так и с позиции аналитика.

«Chief Digital Officer понимает и использует данные бизнес-аналитики, повышая уровень знаний компаний о психологии пользователей и поведении клиентов, — говорит Оливер Наими, старший директор глобальной интернет-платформы и аналитик корпорации Sony[1]. — Аналитические данные, полученные из цифровых каналов, пока остаются новой парадигмой, так что задача определить правильные показатели может оказаться непростой. Директор по цифровым технологиям может повлиять на повышение эффективности компании благодаря обеспечению действенных аналитических выводов на основе измерения, анализа и оптимизации данных бизнес-аналитики, полученных от всех цифровых инициатив через разные каналы».

Заключение

Надеюсь, теперь разница между этими двумя важными функциональными позициями стала ясна. Как видите, для создания в компании корпоративной культуры на основе данных нужно, чтобы эта идея нашла поддержку у высшего руководства и активно продвигалась по всей вертикали в компании. Руководитель должен быть сконцентрирован на масштабном ви́дении того, чего может и должна достигнуть компания с помощью данных, информации и аналитики. Именно руководителю следует фокусироваться на данных и аналитической стратегии в поисках новых возможностей, определять ключевые показатели и при необходимости реорганизовывать структуру компании, чтобы максимизировать гибкость, продуктивность и эффективность.

Как понять, нужно ли в компании создать какую-то из этих позиций или обе сразу? Как именно будет называться эта должность, не столь важно, главное, чтобы *кто-то* взял на себя эту стратегическую роль[2].

---

[1] URL: http://www.oliviernaimi.com/the-emerging-chief-digital-officer.html.

[2] Franks B. Do You Know Who Owns Analytics at Your Company?, Harvard Business Review, September 23, 2014. URL: https://hbr.org/2014/09/do-you-know-who-owns-analytics-at-your-company.

Для начала отметим, что не так уж много компаний, в которых есть две эти должности одновременно. В недавнем отчете McKinsey[1] упоминается по крайней мере одна неназванная «крупная финансовая консультационная компания», которая ввела должность CDO. «CDO находится в прямом подчинении у CIO, но ежедневно работает с директором по аналитике, чтобы помочь объединить данные и новые аналитические инструменты для ускорения изменения процессов работы с клиентами». Однако это скорее исключение. Как правило, наличие двух этих должностей создает некоторую путаницу и кажется избыточным.

Традиционно должность CDO чаще вводится в компаниях, действующих в областях с жестким регулированием. Вероятно, эта тенденция продолжится, так как другие компании в этих областях стремятся скопировать этот подход. Тем не менее пока не сложилось единого мнения, насколько эта тенденция долгосрочна. Исключение составляет только сектор государственного и муниципального управления, который в большей мере сконцентрирован на прозрачности и открытости: в этом случае у CDO более четкие перспективы на будущее и преимущество перед CAO.

Как уже упоминалось, должность CAO появляется, как правило, в тех компаниях, где уже ведется аналитическая работа и требуется расширить, усилить и популяризировать это направление. Если в вашей компании сложилась подобная ситуация, это имеет смысл. В целом, если сомневаетесь, я бы рекомендовал остановить выбор на позиции CAO, так как данные быстрее имеют шанс стать общедоступными, кроме того, продвигать ценность аналитики в компании будет несколько проще.

Как бы ни назывались эти руководители, чтобы эффективно выполнять возложенные на них задачи, они должны тесно взаимодействовать с другими топ-менеджерами компании — генеральным директором и советом директоров — и получать от них поддержку. Они должны располагать бюджетом, командой, а также возможностью пробиться через границы отдельных бизнес-подразделений и создать открытую корпоративную культуру, стимулирующую обмен данными для формирования более богатого и ценного контекста. В итоге это создаст среду, в которой будут процветать аналитика, ее выводы и влияние данных.

[1] Brown B., Court D. and Willmott P. Mobilizing your C-suite for big-data analytics, McKinsey Quarterly, November 2013. URL: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/mobilizing-your-c-suite-for-big-data-analytics.

У индивидуальных источников данных продолжительный жизненный цикл, к тому же их можно использовать для разных продуктов, анализа, проектов. К данным следует относиться «как к программе, а не как к проекту». Это означает, что стоит оторваться от ИТ и подойти к этому вопросу с позиции бизнеса. Опять-таки это больше соотносится с позицией директора по аналитике, но лучше иметь CDO, который подчиняется CTO, чем не иметь руководителя в области данных вообще. Следующая глава будет посвящена изучению чрезвычайно важного и широко обсуждаемого аспекта работы с данными — конфиденциальности информации (или ее отсутствию) и этике. Как компания с корпоративной культурой на основе данных должна обращаться с персональной информацией?

ГЛАВА 12

Вопросы конфиденциальности, этики и риска

У вас в любом случае нет никакой конфиденциальности. Смиритесь.

Скотт Макнили[1]

Человеку, подчиняющемуся нормам морали, следует делать чуть больше, чем от него требуется, и чуть меньше, чем ему разрешено.

Майкл Джозефсон

В предыдущей главе я цитировал Патиля и Мейсон, которые утверждали: «У каждого сотрудника компании должен быть доступ к такому количеству данных, которое только возможно на законных основаниях». Теоретически я с этим согласен, но на практике возникают очень важные моменты, связанные с конфиденциальностью, этикой и безопасностью, которые следует принимать во внимание. В большинстве случаев такие вопросы, как кто и к каким данным должен иметь доступ или как можно *использовать* полученные данные, больше относятся к области этических норм, которых придерживается сам сотрудник, чем к области, которую регулирует законодательство. В корпоративной культуре на основе данных принято уважать как силу данных, так и природу людей, которые становятся источниками этих данных.

Как компания, в которой развито управление на основе данных, должна работать с данными своих пользователей или клиентов с точки зрения этих трех перспектив?

[1] Sprenger P. Sun on Privacy: «Get Over It», Wired, January 26, 1999. URL: http://archive.wired.com/politics/law/news/1999/01/17538.

Я исхожу из предположения, что у компании с управлением на основе данных:

— больше объем данных;
— более обширный контекст, чем у других компаний;
— больше точек интеграции между неразрозненными источниками информации;
— лучше доступ к данным и прозрачность;
— больше сотрудников в компании обладает навыками аналитической работы;
— больше аналитиков, способных замечать неявные закономерности.

ПРИНЦИПЫ КОНФИДЕНЦИАЛЬНОСТИ

В 1998 году Federal Trade Commission (Федеральная комиссия по торговле США) опубликовала важный документ под названием «Защита личной информации онлайн: доклад для Конгресса» («Privacy Online: a Report to Congress»)[1]. Сегодня большинство содержащейся в нем информации кажется устаревшей. Например, на тот момент только 14% детей всех возрастов пользовались интернетом. Сегодня 80% детей в возрасте до пяти лет пользуются Всемирной паутиной еженедельно[2]. Тем не менее один аспект выдержал проверку временем — это пять основных принципов защиты личной информации.

Уведомление/осознанность

«Пользователи должны быть уведомлены о политике использования данных конкретной компанией, прежде чем у них начнут собирать персональные данные».

Выбор/согласие

«Пользователям должны быть предложены варианты, как могут быть использованы их персональные данные».

Доступ/участие

«У пользователей должна быть возможность доступа к своим персональным данным, то есть возможность увидеть, как их данные отражаются в системе хранения данных компании, а также возможность подтвердить точность и полноту данных».

---

[1] Federal Trade Commission. Privacy Online: A Report to Congress, June 1998. URL: https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf.

[2] URL: http://content.usatoday.com/communities/technologylive/post/2011/03/study-80-percent-of-children-under-5-use-internet-weekly/1#.WOYFZLvyi2x

Полнота/безопасность

> «Управленческие и технические способы защиты против утери данных, а также разрешенный доступ, удаление, использование или обнародование данных».

Исполнение/корректировка

> Механизм исполнения других принципов.

Иными словами, по моему мнению, больше данных, больше доступа, больше аналитики означают бóльшую власть и больше риска.

Данные могут наделить властью, но также могут быть очень опасными. Поэтому в этой главе мы рассмотрим некоторые вопросы конфиденциальности, этики и риска, коснемся некоторых опасностей и внешне не совпадающих интересов компаний и их пользователей. По моему мнению, основополагающий принцип, которого компании должны придерживаться в своей деятельности, — эмпатия. Руководствуясь нормами морали и этики как на уровне общей политики компании, так и в обучении сотрудников этическому поведению, и ставя интересы пользователей выше всего, компания сможет завоевать и сохранить доверие своих пользователей, защитить интересы — свои и своих пользователей — и таким образом снизить некоторые из рисков.

Уважайте конфиденциальность

К тому моменту, когда автомобиль подъехал к офису компании Uber на Лонг-Айленде, Джош Морер, управляющий подразделением Uber в Нью-Йорке, уже стоял на ступеньках здания с айфоном в руках. Когда журналистка Джоана Буйян вышла из автомобиля, Джош сказал: «А вот и вы. Я отслеживал ваш путь»[1]. Он использовал корпоративный инструмент под названием God View, который предположительно доступен большинству сотрудников Uber и обеспечивает наблюдение за автомобилем и местоположением клиента в режиме реального времени. Это был не первый раз, когда компания Uber нарушала конфиденциальность пользователей. На вечеринке по поводу открытия офиса в Чикаго три года назад участники в режиме реального времени наблюдали за передвижениями по Нью-Йорку пользователей, личнос-

[1] Kosoff M. Uber's Top New York Executive Is Being Investigated After Using Uber's «God View» Tool To Track A Journalist's Location (http://bit.ly/bi-uber-godview), Business Insider, November 19, 2014. URL: http://www.businessinsider.com/ubers-new-york-manager-investigated-for-using-god-view-2014-11.

ти которых можно было легко идентифицировать, в том числе венчурного капиталиста Питера Симса[1].

Все дело в том, что ни в одном из случаев клиенты не были оповещены о том, что данные о них будут использоваться подобным образом, и не давали согласия на это. Да, возможно, компании Uber требуется такой доступ и инструменты для повышения качества обслуживания клиентов, но этот подход выходит за рамки действий, определенных Федеральной комиссией по торговле как «необходимые для исполнения условий договора». В обоих вышеприведенных случаях явно наблюдалось превышение полномочий.

В этих конкретных случаях фактического вреда нанесено не было, но легко можно представить себе сценарий, несущий потенциальную угрозу: человек, скрывающийся от агрессивного партнера; пассажир, вышедший возле клиники, проводящей тестирование на ВИЧ; знаменитость, не желающая встречаться с навязчивым поклонником. (Дана Бойд приводит дополнительные примеры в контексте настроек конфиденциальности Facebook[2].)

В правилах хранения и использования персональной информации, которые фактически выполняют функцию соглашения между пользователем или клиентом и компанией, должно быть четко определено, кто занимается сбором данных, как этот сбор данных осуществляется, каким образом эти данные будут и не будут использоваться, на каких условиях доступ к ним могут получить третьи лица, каковы последствия отказа предоставить согласие, а также «меры, предпринятые стороной, осуществляющей сбор данных, для обеспечения конфиденциальности, полноты и качества данных».

Очевидно, что компания Uber нарушила эту политику конфиденциальности[3], однако ее соблюдение — это не единственный вопрос, на котором должны сконцентрироваться все компании. Пользователи обязаны понимать условия политики безопасности. Часто лицензионные соглашения с конечными пользователями (EULA) бывают очень длинными. Представьте: объем «Гамлета» — 30 тыс. слов, а пользовательское соглашение Paypal[4] — 50 тыс., что приблизительно эквивалентно первым семи главам нашей книги. Эти документы содержат

[1] Sims P. Can We Trust Uber? URL: http://bit.ly/sims-uber) Silicon Guild, September 26, 2014.

[2] URL: http://www.danah.org/papers/talks/2010/SXSW2010.html.

[3] См. правила хранения персональных данных Uber's Data Privacy Policy. URL: http://bit.ly/uber-privacy-policy и статью Слейта — URL: http://bit.ly/slate-uber-privacy.

[4] URL: http://www.bbc.com/news/technology-22772321.

кучу юридических терминов, но «простые пользователи» должны согласиться со всеми пунктами. Любой компании стоило бы проявить уважение к своим пользователям и сформулировать политику конфиденциальности таким языком, чтобы она была понятна всем пользователям (то есть была удобочитаема для человека). (Любые порочащие измышления, что юристы лишены человеческих качеств, случайны.) Замечательный пример мирного сосуществования юридических терминов и доступности восприятия для обычного человека — политика конфиденциальности популярной онлайновой платформы CodePen[1].

Если я просто шучу по поводу лицензий, понятных для обычных пользователей, то для компании Creative Commons[2] это стало важным отличием: лицензии и правовые инструменты этой организации имеют «трехслойный» дизайн, чтобы сделать защиту «эффективной, юридически осуществимой и незаметной».

Текст, понятный для пользователей

Пользователи должны быть в состоянии понять, с чем они соглашаются. Социальная сеть Facebook, которая долгие годы буквально утопала в спорах и претензиях по поводу настроек конфиденциальности, в последнее время сделала значительные шаги по улучшению ситуации: ее правила хранения и использования персональных данных по-прежнему очень длинные, но теперь гораздо более четко структурированы и доступны для понимания пользователям, не имеющим юридического образования[3].

Юридический текст

Традиционный правовой инструмент, текст, написанный на «юридическом» языке, обеспечивающий всестороннюю защиту.

Версия, «читаемая машиной»

Применение технологических подходов, например P3P или Creative Commons[4], делает тексты лицензий доступными для понимания системами ПО, поисковыми системами и другими видами технологий[5].

---

[1] URL: https://blog.codepen.io/legal/privacy/.

[2] URL: https://creativecommons.org/.

[3] URL: https://www.facebook.com/policy.php.

[4] URL: https://creativecommons.org/.

[5] Лоуренс Лессиг видит это следующим образом: https://www.youtube.com/watch?v=cXoXXbo_mL4.

Итак, уважайте своих пользователей, предлагая им правила политики конфиденциальности, которые они могут понять и по поводу которых могут принять информированное решение. Уважайте конфиденциальность пользователей, строго придерживаясь принципов и условий, прописанных в вашем соглашении.

НЕПРЕДНАМЕРЕННАЯ УТЕЧКА ИНФОРМАЦИИ

Случай на вечеринке по поводу открытия офиса Uber — пример того, как данные пользователей или контекст (кто и где находился, в какое время) попали в открытый доступ. При этом по мере того как все больше компаний внедряют управление на основе данных, я наблюдаю все больше случаев, как компании собирают множество на первый взгляд безобидных сведений, но чем большей статистической значимостью они обладают, тем серьезнее риск их непреднамеренной утечки.

Несколько лет назад, как раз в разгар скандальных откровений Эдварда Сноудена[1] по поводу несанкционированной слежки АНБ США и горячих дебатов относительно конфиденциальности, я опробовал инструмент под названием immersion[2] («погружение»)[3]. Этот инструмент анализировал только *метаданные* сообщений электронной почты. Метаданные — характеристики сообщения: отправитель, получатель, время отправления. При этом анализ содержания сообщения не проводится. Может показаться, что у этих метаданных весьма ограниченный спектр применения. Однако, когда я воспользовался этим инструментом относительно своей учетной записи электронной почты, я был поражен. Этот инструмент наглядно показал мне группы людей из разных сфер моей жизни, которые знали друг друга, которые могли представить меня другим людям, а также относительную силу этих социальных связей. Фактически это было весьма точным отражением моей социальной сети на тот момент. И это без доступа к содержанию сообщений. В другом примере Латания Суини показывает, что можно идентифицировать 87% американцев исключительно по информации

---

[1] Эдвард Джозеф Сноуден (Edward Joseph Snowden, p. 1983) — американский технический специалист и спецагент, бывший сотрудник ЦРУ и Агентства национальной безопасности США. *Прим. ред.*

[2] URL: https://immersion.media.mit.edu/.

[3] См. также: Chen B. X. Using E-Mail Data to Connect the Dots of Your Life, The New York Times, July 5, 2013. URL: https://bits.blogs.nytimes.com/2013/07/05/using-e-mail-data-to-connect-the-dots-of-your-life/.

о почтовом индексе, поле и дате рождения[1]. У нас все больше данных и все более сложные инструменты и навыки, позволяющие нарисовать общую картину. Это можно сравнить с картиной Жоржа Сёра[2], выполненной в манере пуантилизма[3], только данными.

Незначительные сведения из нашей онлайн-активности и реальной жизни дополняют картину, и аналитики всегда бывают счастливы собрать все кусочки воедино. Однако делать это следует, не преступая этические нормы, которые преимущественно не закреплены законодательно и определяются тем, как их воспринимает сам аналитик.

Один из примеров использования конфиденциальной информации, когда все происходило в рамках закона, но привело к неблагоприятным последствиям, связан с компанией Target. В статье[4], опубликованной в New York Times и вызвавшей оживленное обсуждение среди специалистов по работе с данными, журналист Чарльз Дахигг рассказывает, как специалисты по маркетингу компании Target попросили одного из аналитиков компании, Эндрю Пола, определить группу покупательниц, которые были беременны, чтобы знать эту информацию до того, как появятся официальные сведения о рождении ребенка. Маркетологи предположили, что, если выделить эту категорию женщин достаточно рано, есть больше шансов заинтересовать их купонами и создать базу лояльных клиентов.

Эндрю и его коллегам удалось успешно определить, какие покупки совершали беременные женщины, и компания начала таргетированную рассылку купонов. Все это вполне в рамках закона, но причина, по которой эта история вызвала такой резонанс среди аналитиков, касается этической стороны и истории отца одной из беременных девушек.

Компания Target занимается рассылкой персонализированных буклетов с купонами. Как правило, потребители охотно пользуются купонами на те товары, которые они в любом случае покупают. Однако реакция беременных женщин была негативной. Поэтому компания

---

[1] Sweeney L. Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, 2000. URL: http://dataprivacylab.org/projects/identifiability/paper1.pdf.

[2] Жорж-Пьер Сёра (Georges Seurat, 1859–1891) — французский художник-постимпрессионист, основатель неоимпрессионизма, создатель метода живописи под названием дивизионизм, или пуантилизм. *Прим. перев.*

[3] Пуантилизм, или дивизионизм — стилистическое направление в живописи неоимпрессионизма, возникшее во Франции около 1885 года, в основе которого лежит манера письма раздельными мазками правильной, точечной или прямоугольной формы. *Прим. перев.*

[4] Duhigg C. How Companies Learn Your Secrets, The New York Times, February 16, 2012. URL: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

начала добавлять купоны на товары, не связанные с беременностью, например купон на покупку газонокосилки вместе с купоном на покупку подгузников, чтобы замаскировать то, что они знали о своих покупательницах. Вот что рассказывает один из руководителей компании: «Мы обнаружили: если женщина не считала, что за ней шпионят, то спокойно использовала купоны. То есть она просто была уверена, что все остальные жители ее квартала получают точно такие же рассылки с купонами на подгузники и детские кроватки. Если покупательницу не спугнуть, наша стратегия работает».

Компания прилагала все усилия, чтобы замаскировать информацию, известную им о своих покупательницах, но это не ускользнуло от внимания одного неравнодушного отца:

> Примерно через год после того, как Пол разработал свою прогнозную модель, в офис компании Target в Миннеаполисе вошел мужчина и потребовал встречи с менеджером. Как рассказал один из сотрудников компании, присутствовавший при разговоре, мужчина, сжимавший в руке пачку купонов, был в бешенстве.
>
> «Моя дочь получила это по почте, — заявил он. — Она еще учится в старшей школе, а вы посылаете ей купоны на покупку детской одежды и кроватки. Вы что, занимаетесь пропагандой подростковой беременности?»
>
> Менеджер понятия не имел, о чем говорит этот мужчина. Он взглянул на буклет. Никаких сомнений: буклет был адресован дочери этого мужчины и содержал рекламу детской одежды и мебели, а еще фотографии розовощеких младенцев. Менеджер принес свои извинения, а затем позвонил через несколько дней, чтобы извиниться еще раз. Отец на другом конце провода был явно смущен. «Я поговорил с дочерью, — объяснил он. — Кажется, в моем доме происходит нечто, о чем я не имел ни малейшего представления. Она должна родить в августе. Это я должен принести вам извинения».

Эта рекомендация товаров в форме купонов выдала семье девушки ту информацию, которую она от них скрывала. Это была утечка не P.I.I. — данных, обеспечивающих идентификацию личности, — а, как метко выразилась Дана Бойд, P.E.I. — данных, *ставящих в неловкое положение*.

Большинство медицинских данных попадает под защиту, например, Закона США о сохранении медицинского страхования и персонифицированном учете в здравоохранении (HIPAA) 1996 года. В данном случае вывод об «интересном» положении девушки был сделан на основе информации о невинных товарах, которые она покупала ранее, например таких, как лосьон без запаха. С правильными данными и инструментами аналитики обладают практически безграничными возможностями вмешиваться в чужие жизни, поэтому им следует тщательно просчитывать возможные последствия этого вмешательства, не только для того, чтобы «не спугнуть» людей.

Практикуйте эмпатию

По моему убеждению, компании с управлением на основе данных должны уважать права и чувства своих пользователей. Возможно, эти компании стремятся постоянно выходить за рамки и собирать все больше и больше данных, способных обеспечить им «пищу» для рекламных кампаний, сервисов и продуктов на основе данных, но в долгосрочной перспективе им гораздо выгоднее завоевывать и поддерживать доверие пользователей.

Самый простой тест, когда вы выбираете новые настройки конфиденциальности или разрабатываете новые стратегии, характеристики или кампании, связанные с данными: вам понравится пользоваться этим самому или предложите вы это своим близким друзьям? Если нет, откажитесь от этой идеи.

В компании Warby Parker главный юридический консультант Анджали Кумар даже дала этому название — фактор «фу». Это качественный показатель меры, как «не спугнуть»; естественно, он не закреплен законодательно, но это напоминание о том, что мы подчиняемся не только юридическому закону, но и «законам совести»: ставим себя на место потребителя и проявляем эмпатию. Как бы себя чувствовал покупатель?

Приведу пример: однажды Анджали возвращалась в Нью-Йорк на поезде. Ее попутчик сошел на одной из станций, но, к сожалению, забыл свои очки. Оказалось, что это очки от компании Warby Parker. Когда Анджали пришла в офис, у нас с ней состоялось серьезное обсуждение, насколько корректно мы поступим, если попробуем найти этого мужчину и вернуть ему очки. Какой в этом фактор «фу»? После долгих размышлений мы решили, что действуем в лучших интересах клиента нашей компании. Мы воспользовались базой данных наших покупателей, чтобы определить того, кто мог быть нам потенциально интересен

(как вы помните, у нас была его оправа, мы знали его пол, примерный возраст и на какой станции он вышел). Затем мы сузили круг, и финальной проверкой для нас стало его фото в социальной сети LinkedIn. Анджали отправила своему забывчивому попутчику новую пару очков, роман Джека Керуака «В дороге» и записку:

> Привет, Майкл! Это может показаться вам странным, но несколько недель назад вы сидели напротив меня в поезде из Бостона в Нью-Йорк и забыли свои очки. По счастливой случайности я работаю главным юрисконсультом компании Warby Parker и просто обожаю хорошие детективные истории… Надеюсь, у вас все в порядке! Кстати, мы заметили, что линзы на вашей паре очков поцарапались, так что решили прислать вам новую пару. Искренне ваша, АК[1].

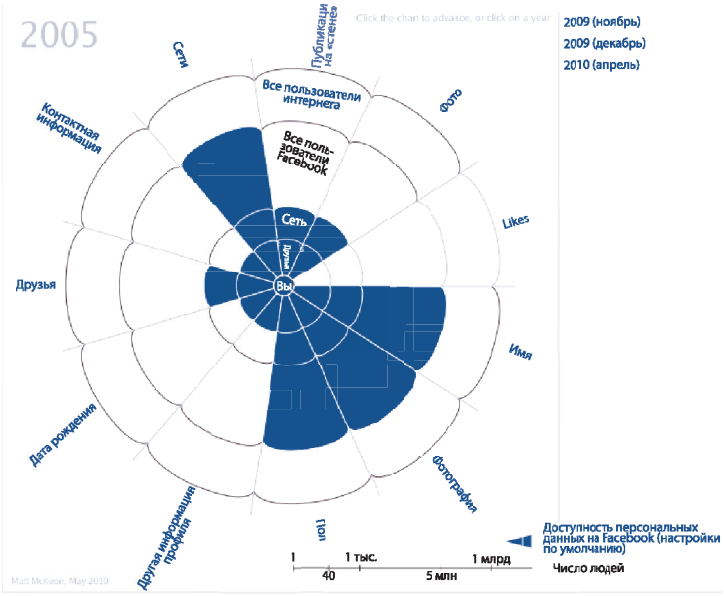
Дело в том, что мы очень серьезно подошли к вопросу использования конфиденциальной информации и поступили так не ради развлечения или потому что у нас была такая возможность. Мы проконсультировались с руководством, насколько корректным будет наш поступок, не напугает ли он нашего клиента и воспримет ли он нашу мотивацию правильно: обеспечить лучшее обслуживание для наших покупателей.

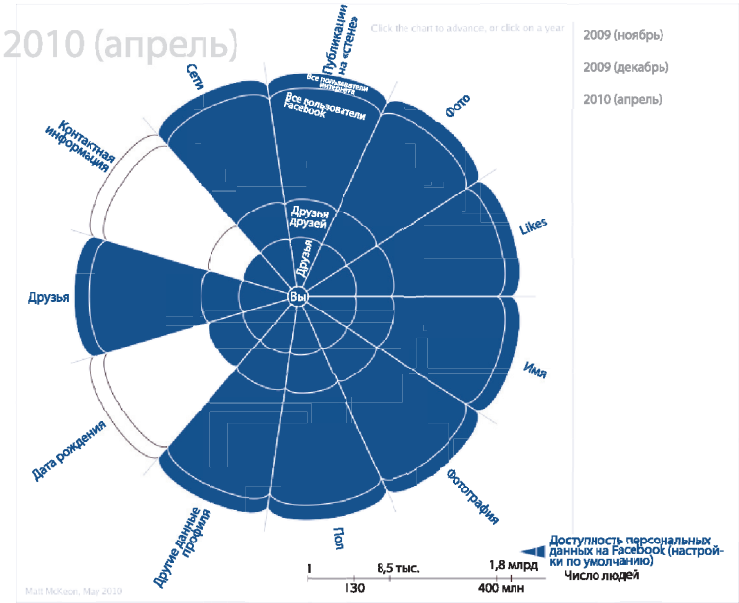
Это был осознанный риск, но, к счастью, клиент оценил наше внимание и написал в социальной сети: «Это лучшее обслуживание, с которым я сталкивался в своей жизни». (Чтобы прояснить ситуацию: мы никоим образом не рекламировали то, что сделали. Единственной нашей мотивацией была польза для клиента. В прессу эта история просочилась, потому что наш чрезвычайно довольный клиент рассказал обо всем на своей страничке социальной сети, а журналист, опубликовавший впоследствии статью, входил в список его контактов.)

ВЫХОДЯ ЗА РАМКИ

Социальная сеть Facebook постоянно испытывает разногласия со своими пользователями, часто выходя за рамки того, какой информацией можно делиться и с кем, а в нескольких случаях даже была вынуждена уступить, когда жалобы от пользователей начали поступать в особо больших количествах. По заявлению Марка Цукерберга, защита персональных данных — «вектор, вокруг которого строится деятельность Facebook», а сам он уверен, что Facebook просто следит

[1] Phelps S. Heroic Customer Service by a Senior Executive at Warby Parker. Forbes, August 1, 2014. URL: https://www.forbes.com/sites/stanphelps/2014/08/01/heroic-customer-service-by-a-senior-executive-at-warby-parker/.





Источник: The Evolution of Privacy on Facebook (http://mattmckeon.com/facebook-privacy/)

за изменением социальных норм: «Теперь люди чувствуют себя гораздо комфортнее, когда открыто делятся самой разной информацией с большим количеством других людей. Эта социальная норма просто изменилась со временем».

Изменения в этом вопросе — в настройках конфиденциальности по умолчанию для различных аспектов на сайте — просто поразительны. Сравните следующие два графика. Первый показывает настройки по умолчанию в 2005 году, а второй — те же самые настройки через пять лет в 2010 году.

Компании с управлением на основе данных обладают огромной властью. Применяйте ее во благо.

ПРЕДОСТАВЬТЕ ВЫБОР

По возможности предоставьте пользователям интуитивно понятные, подходящие инструменты контроля над тем, как используются их данные или каким образом они доступны остальным. Например, это может быть возможность контролировать тип или частоту маркетинговых рассылок, возможность отказываться от принудительных уведомлений от приложений и предложений партнерских организаций. Больше противоречий вызывает то, что персональные данные могут передаваться третьим лицам. Именно это стало источником проблем для разных социальных сетей (Facebook — лишь один пример, см. врезку выше), где изменение настроек по умолчанию еще хуже сказывается на защите персональных данных.

Одна из проблем в том, что даже когда компания действительно обеспечивает защиту персональной информации, многие пользователи не понимают, какие варианты для них доступны. В итоге у большинства из них так и остаются настройки по умолчанию. В этом случае у компании есть по крайней мере два способа действий. Во-первых, поставить себя на место пользователя: сделать меры контроля простыми, интуитивно понятными и четко задокументированными. Во-вторых, поставить защиту персональной информации и уважение во главу угла и действовать исходя из того, что клиент соглашается на использование информации. Обеспечьте пользователям выбор и возможность контроля.

Компания Netflix предлагает интересную возможность в панели настроек пользователя. Пользователь может отказаться от участия в A/B-тестировании (рис. 12.1). Я никогда не видел подобного у других сервисов.

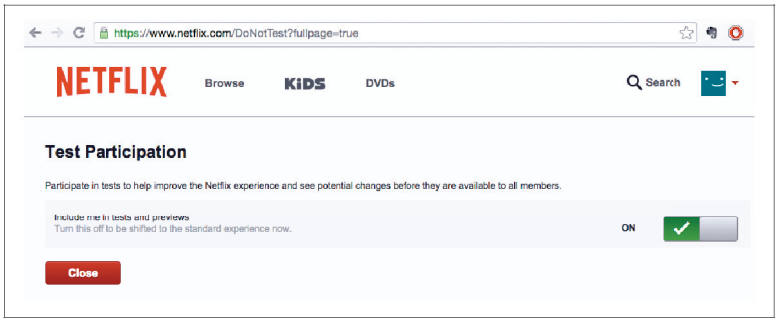


Рис. 12.1. Netflix (https://www.netflix.com/ru/) предлагает пользователям отказаться от участия в A/B-тестах в настройках своей учетной записи

Здесь налицо конфликт интересов. Компания поступает справедливо, предоставляя выбор пользователям. При этом Netflix активно проводит A/B-тестирования. Чтобы как можно быстрее получить данные A/B тестов, на основе которых можно сделать обоснованные заключения, требуется большая выборка. Отказ пользователей от участия в A/B тестах уменьшает размер выборки, увеличивает время проведения тестов и, возможно, влияет на объективность выборки.

Однако смею выдвинуть предположение, что только очень малая часть пользователей применила эту опцию. Если я прав, то подписчики только выиграли от этого (они могут отказаться от участия в экспериментах, если у них возникли сомнения), а низкий уровень отказа практически не влияет на результаты тестирования и на компанию в целом. В этой ситуации компания Netflix заработала себе хорошую репутацию и почти ничего не потеряла. В этом с нее можно брать пример.

Качество данных

Один из основных принципов защиты персональных данных Федеральной комиссии по торговле — доступ/участие, то есть возможность для пользователя видеть, какая информация о нем хранится в базе данных организации, и возможность подтвердить ее или исправить.

На мой взгляд, это, вероятно, один из наименее проработанных из пяти принципов. Большинство онлайн-сервисов обеспечивают пользователям возможность редактировать информацию профиля и обновлять данные об адресе пользователя, адресе его электронной почты и другую идентифицирующую пользователя информацию. Некоторые

организации, особенно социальные сети, позволяют экспортировать архивы данных (например, Twitter и Facebook). Что в большинстве случаев сделать невозможно, так это отредактировать все предшествующие данные, например предыдущие заказы, или просмотреть все «сопутствующие» данные, которые организация о вас собрала (например, из переписи населения США, единой базы недвижимости, от компаний, торгующих данными, из социальных сетей и так далее). Откровенно говоря, это сложно обеспечить. Кроме того, пользователям было бы сложно понять разрозненные записи баз данных. Это могло бы нарушить соглашения относительно данных, приобретенных у других организаций, и, возможно, выдало бы некоторые секреты внутренней кухни компании. Так что я не наблюдаю значительного прогресса в этой области.

Хотя компании с управлением на основе данных, конечно, должны сделать максимально простым процесс обзора и исправления основной информации о пользователях. Это отвечает интересам как пользователей, так и компаний. При наличии данных из разных внутренних источников, например из заявки на кредит и информации по текущему счету в том же банке, есть вероятность привязать одного клиента к идентифицирующей информации другого клиента или внести небольшие изменения в данные на разных этапах ввода (например, «улица» вместо «ул.» или «кв. 6» вместо «№ 6»). Чем проще будет исправить и стандартизировать данные о пользователях, тем эффективнее окажется работа компании на основе данных.

Если бы вы увидели мою учетную запись в Netflix, то получили бы крайне приблизительное представление о моих предпочтениях. Вы увидели бы рекомендации относительно очень разных телесериалов, таких как The Magic School Bus, Gilmore Girls и M*A*S*H[1]. Это создает не совсем верное представление о том, что смотрю лично я. Все дело в том, что этой учетной записью пользуются все члены моей семьи, а потому просмотры и последующие рекомендации фактически сделаны для нас четверых, а не для меня одного. И если у компании Netflix есть концепция профиля, которая помогает выделить таких множественных пользователей, эта функция недоступна на устройстве, с которого я пользуюсь этим сервисом.

[1] The Magic School Bus («Волшебный школьный автобус») — познавательно-приключенческий мультсериал по мотивам комиксов Джоанны Коул; Gilmore Girls («Девочки Гилмор») — американский комедийно-драматический телесериал; M*A*S*H — американский телесериал, созданный по мотивам романа Ричарда Хукера «МЭШ: роман о трех армейских докторах», последующей серии рассказов и кинофильма M*A*S*H. *Прим. перев.*

Обеспечьте пользователям возможность предложить дополнительный контекст относительно своих данных, который сможет оказать влияние на то, как компания оценивает или использует эту информацию. Например, интернет-магазин Amazon предлагает функцию «Улучшить рекомендации» (Improve Your Recommendations), где пользователь может указать, что какой-то из товаров он приобретал в подарок или что товар не следует использовать при формировании рекомендаций. Пользователь может не хотеть, чтобы какой-то товар использовался при формировании рекомендаций и чтобы ему показывали список похожих товаров в будущем, по многим причинам, в том числе потому что это может поставить его в неловкое положение. Тем не менее, какими бы ни были эти причины, предлагая пользователю возможность исправить, отфильтровать или исключить какую-то информацию, компания получает более точное представление о намерениях пользователя, контексте или его предпочтениях. Этот принцип действует и в обратном направлении: возможно, пользователь почувствует себя более уверенно, если получит информацию, почему ему была предложена подобная рекомендация. Например, в своей учетной записи Netflix я недавно увидел рекомендацию обратить внимание на телесериал «Частный детектив Магнум», «потому что вы смотрели M*A*S*H». Эта рекомендация имеет смысл. Такое объяснение также сможет выявить неточную информацию, которую пользователь хотел бы исключить или исправить.

Итак, благодаря добавлению подобных функций компания может стимулировать двусторонний диалог между собой и пользователем, что приведет к получению более точных данных и контекста и, в конце концов, к предоставлению пользователям более качественного сервиса.

Безопасность

Ранее я упоминал, что меры по снижению риска часто способны ограничить деятельность гораздо больше, чем требуется законодательно. Почему так происходит?

Начнем с простого примера. У многих специалистов по работе с данными, например технических специалистов и администраторов баз данных, имеется доступ к сырым данным о пользователях. Эти данные могут включать имя, адрес, номер телефона, электронную почту и другую информацию, идентифицирующую человека. Закон это разрешает. Такой доступ им предоставляется потому, что они выполняют свои функциональные обязанности, обеспечивая правильный сбор

и хранение данных, чтобы организация могла выполнять свои обязательства по деловым сделкам.

Теперь представим специалиста по анализу, который должен проанализировать количество проданных единиц товара в разные дни. Законодательно ничего не мешает этому аналитику получить доступ к сырым данным о пользователях. Однако действительно ли ему требуется такой уровень детализации? Требуется ли ему доступ к этим данным для проведения своего анализа? Фактически ему не обязательно знать, что набор садовой мебели заказала именно Белинда Смит, проживающая по такому-то адресу, с таким-то номером телефона и адресом электронной почты. Все, что нужно знать этому аналитику, — то, что торговая единица 123456 была продана в определенный день.

В большинстве случаев при анализе данные агрегируются, и информация, идентифицирующая пользователей, не требуется.

В своей книге Dataclysm сооснователь сервиса для знакомств OKCupid Кристиан Раддер представляет ряд примеров анализа на основе данных с сайта. За исключением данных медицинского характера вы вряд ли найдете где-то более точную информацию о пользователях, чем на сайте знакомств. В профилях посетителей сайта есть фотографии, указан пол, возраст, сексуальные предпочтения, сферы интересов и другая очень личная информация. Кристиан Раддер рассказывает (с. 233), как он работал с данными:

> Любой тип анализа проводился анонимно, а данные агрегировались. Я очень внимательно отнесся к исходным данным. Ни в одних данных не содержалось информации, идентифицирующей пользователя… Там, где использовалась персональная информация, данные шифровались. Кроме того, при любом типе анализа объем данных был ограничен только до необходимых переменных, так что не было никакой возможности связать что-то с конкретными людьми.

Все эти меры предосторожности Кристиан предпринимал по нескольким причинам. Во-первых, он не хотел, чтобы какая-то информация повлияла на объективность результатов анализа. Любой аналитик стремится к тому, чтобы результаты его анализа были максимально объективными. Дополнительная информация может исказить интерпретацию. Например, если вы увидите, что имя пользователя Гертруда,

как вам кажется, она молодая или старая? Старая, верно?[1] Эти предположения формируются у вас неосознанно. Вы снизите риск неосознанных предположений, отказавшись от включения дополнительных переменных, и повысите шанс обнаружения истинных закономерностей в агрегированных данных.

Во-вторых, аналитики часто копируют данные для проведения анализа и разработки моделей с помощью других инструментов. Так что иногда, когда один аналитик пользуется инструментом бизнес-аналитики для агрегирования данных, другому аналитику может быть необходимо обработать эти данные в Python или R для разработки сложных прогностических моделей. Часто это означает необходимость экспортирования данных из основного источника хранения данных в файлы на ноутбуке. Каждая копия помимо основного источника данных увеличивает риск для компании. Ноутбук можно украсть или взломать. Аналитик, работающий на своем ноутбуке в зале аэропорта или в кафе Starbucks, подвергается риску, что кто-то увидит информацию на мониторе. Так что чем меньше информации он хранит таким образом и чем больше уровней защиты, тем лучше.

Именно по этим причинам многие компании предпочитают обезличивать информацию, которая отображается в базах данных и инструментах бизнес-анализа для составления отчетов и проведения анализа. Имена, адреса, адреса электронной почты полностью скрываются или зашифровываются.

Например, адрес электронной почты belinda.smith@example.com с помощью хеша SHA-256 можно зашифровать как `f7bf49636a69c6ed45da8dc8d3f445a8a5e6bcc2e08c9a6b2bb66446c402f75c`.

(Это действует в одном направлении: можно очень просто превратить адрес электронной почты в зашифрованную последовательность символов, но крайне сложно, если возможно вообще, выделить адрес электронной почты из этой последовательности.). Опять-таки, в большинстве случаев законодательно компании не обязаны это делать, но это явно имеет смысл.

Чем больше количество копий, тем выше риск. Чем больше количество файлов для чтения человеком, тем выше риск. Чем больше передвижений и интеграций разных источников данных — что характерно для компании с управлением на основе данных, в которой продвигается

[1] URL: https://www.google.ru/search?q=gertrude&tbm=isch&gws_rd=cr&ei=yKOwWL6oNKKR6ATMgLSIDg .

обмен информацией, — тем выше риск. Треть руководителей[1] признались, что «в их компании не удается внедрить управление на основе данных частично из-за вопросов конфиденциальности и безопасности, которые возникают при обмене информацией».

Мы можем сделать заключение в виде принципов, перечисленных ниже.

— Каждый сотрудник, которому требуется доступ к данным для выполнения своих профессиональных обязанностей, имеет этот доступ.
— Каждый сотрудник имеет доступ только к тем данным, которые требуются ему для выполнения профессиональных обязанностей.
— К персональной информации, такой как данные о пользователях и рекомендации, следует относиться с повышенным вниманием: доступ к ней должен быть максимально ограничен, информация должна быть обезличена и зашифрована.

Обеспечение исполнения

По заявлению Федеральной комиссии по торговле, «согласно общему мнению, основные принципы защиты конфиденциальности могут быть эффективны только в том случае, если присутствует механизм обеспечения их исполнения».

Конечно, сегодня многие нормативные акты регулируют процессы сбора и использования данных, а также вопросы конфиденциальности. В числе примеров Закон о защите личных сведений детей в интернете (COPPA), Закон США о сохранении медицинского страхования и персонифицированном учете в здравоохранении (HIPAA), совместимость со стандартом безопасности PCI при проведении платежей.

Очевидно, все должны подчиняться требованиям закона. Они обозначают верхнюю границу того, что можно делать с данными на законных основаниях. Однако я убежден, что этого недостаточно. Компании с управлением на основе данных должны руководствоваться в своей деятельности более широкими вопросами этики и фактора «фу» и разрабатывать собственные внутренние правила и принципы деятельности. У них должен быть собственный моральный компас, ориентированный

[1] Аналитическое подразделение журнала Economist, Fostering a Data-Driven Culture (London: Economist Intelligence Unit, 2013). URL: https://www.tableau.com/economist-fostering-data-driven-culture.

на данные. Они должны принимать во внимание, ожидает ли пользователь, что его данные будут использоваться именно так, и будет ли он с этим согласен. Аналитику следует время от времени задавать себе вопрос: «Как бы я чувствовал себя на месте пользователя?» Фактически это может несколько ограничить спектр того, как аналитик, возможно, хотел применить имеющиеся в его распоряжении данные. Подобно специалистам по маркетингу компании Target, всегда найдутся люди, стремящиеся выйти за установленные рамки (в конце концов, им требуется выполнять собственные KPI), поэтому необходима корпоративная культура, руководство на основе данных и обучение, чтобы установить рамки приемлемого.

Заключение

В компаниях с активным использованием данных всегда будет наблюдаться некоторое здоровое напряжение между разными командами: так, например, аналитики всегда будут стремиться создавать самые современные продукты с использованием данных, а более консервативные юристы — минимизировать риски для компании. В то время как законодательные ограничения непреложны, существует обширная серая зона, деятельность в которой не нарушает закон, но может вызывать сомнения с морально-этической точки зрения.

Компания должна уважать своих пользователей и разработать руководство, что считать приемлемым и неприемлемым использованием данных. Очевидно, компании нужно установить ограничительную линию для аналитиков, чья работа наиболее тесным образом связана с данными. В компании Warby Parker мы сформулировали, как каждый из наших типов данных (данные клиентов, данные о продажах и так далее) может или не может быть использован при проведении разных видов анализа или маркетинговых мероприятий. Например, в нашем рецепте на очки обычно указывается дата рождения. Мы считаем, что аналитик может воспользоваться этими данными на *агрегированном уровне*, чтобы лучше понять базу данных наших клиентов за счет изучения распределения по критерию возраста. Однако специалисты по маркетингу не могут на основе этой информации на *индивидуальном уровне* выбрать, например, категорию клиентов в возрасте 25–34 лет.

В компаниях с управлением на основе данных существует более широкий доступ к данным, поэтому информацией могут пользоваться в том числе специалисты, которые не связаны непосредственно

с аналитической работой, но у которых доступ к данным определяется их функциональными обязанностями (например, сотруднику службы по работе с клиентами требуется доступ к их данным). Они используют данные для повышения качества работы. Для этих сотрудников должны быть четкие руководства и система обучения, особенно для молодых специалистов. Например, следует четко заявить, что они не могут использовать информацию о клиентах, об их предпочтениях и так далее в рекламных объявлениях или публикациях на Facebook без их согласия или что они не имеют права изучать базы данных без профессиональной необходимости, например в поисках знакомых, знаменитостей, друзей и так далее. Обеспечьте обучение по этим вопросам. Как сказано в комиксах про Человека-паука: «Большая власть подразумевает большую ответственность»[1]. Компании следует активно заниматься вопросами ответственности и перспективы.

[1] URL: https://en.wikipedia.org/wiki/Uncle_Ben.

Заключение

Информация — это новая нефть!
Клайв Хамби, Dunnhumby

Что для компании означает управление на основе данных? Возможно, вы уже поняли, что ответ на этот вопрос заключается не в обладании новейшими технологиями по работе с большими данными и не в команде блестящих специалистов по аналитике. С ними, несомненно, будет легче, но сама концепция управления на основе данных касается не какой-то конкретной вещи. Скорее, как я уже говорил, она охватывает всю аналитическую цепочку ценности и всю структуру компании. Это отражено на рис. 13.1.

В главах 2 и 3 мы обсуждали самый первый слой — сами данные, как собирать правильные данные и как собирать данные правильно. Помимо этого, требуются люди, обладающие нужными навыками, и инструменты. Кроме того, необходимо проводить обучение, чтобы использовать данные максимально эффективно.

Конечно, в первую очередь речь идет об аналитическом подразделении компании, но в компании с управлением на основе данных количество сотрудников, опирающихся в своей работе на данные, выходит далеко за пределы аналитического подразделения.

Как я неоднократно подчеркивал, у меня нет сомнений, что в компании каждый сотрудник вносит свой вклад в общее дело: это совместная ответственность. Основная аналитическая цепочка идет от специалистов по аналитике и их руководителей к руководителям высшего звена, топ-менеджменту компании и совету директоров. Однако в более демократичной с точки зрения работы с данными среде, где, как отметил Кен Рудин, «каждый сотрудник — аналитик», в обязанности всех сотрудников входит, помимо прочего, активное применение доступных данных, инструментов и обучающих программ, чтобы по возможности включать эти данные в свою работу, сообщать о проблемах

с качеством данных, генерировать достойные тестирования гипотезы, подвергать сомнениям необоснованные стратегии, мнения и HiPPO и в целом использовать данные с максимальной эффективностью.

Одной из задач этой книги было прямое обращение к специалистам по аналитике и их руководителям. Роль этих сотрудников часто недооценивают. Часто фокус и обсуждение сосредоточивают на изменениях, которые требуется проводить «сверху вниз», когда фактически специалисты по аналитике играют ключевую роль в формировании аналогичной корпоративной культуры с нижних уровней компании. Для этого им нужно действовать более активно и сделать свою роль в компании более заметной.

Эту идею очень удачно выразил Чарльз Томас, директор по данным компании Wells Fargo:

> Я называю специалистов по аналитике людьми, которые стимулируют действия: выбирайтесь из своих четырех стен, избавляйтесь от репутации «гиков», демонстрируйте всем свои деловые качества, показывайте, как плоды вашей работы сказываются на всей компании. Вам придется приложить дополнительные усилия, чтобы убедиться, что результаты аналитической работы применяются на всех уровнях компании. Заставьте их работать.

Выходите из своей зоны комфорта и стимулируйте изменения!

Рис. 13.1. Обзор компонентов, из которых складывается компания с управлением на основе данных

Источник: на основе концепции Уэйна Экерсона, изложенной в его книге Secrets of Analytical Leaders

Необходимо добиться оптимальной организации аналитического подразделения (глава 4). Обычно это осуществляется на основе объединенной, или гибридной, модели, когда аналитики работают в разных бизнес-подразделениях, но при этом есть централизованное аналитическое подразделение, в задачи которого входит обучение сотрудников, поддержка, разработка единых стандартов, и где у специалистов по аналитике определен четкий карьерный путь. Специалисты этого подразделения должны быть сосредоточены на качестве работы, и по крайней мере несколько из них должны заниматься предсказательной аналитикой и аналитикой на более высоком уровне, например разрабатывать прогнозные модели и меры по оптимизации. Они должны продвигать свои аналитические выводы и рекомендации и убеждать в них людей, принимающих решения (следующий уровень на рис. 13.1). В идеале они должны получать оценку своей работы по фактическому влиянию на эффективность деятельности компании.

Продвижение комплексной аналитической программы подразумевает наличие сильного руководства на основе данных. Его может осуществлять, например, вице-президент, отвечающий за аналитическое направление, или директор по данным. В компаниях из рейтинга Fortune 500 эта роль все чаще отводится CDO или CAO (глава 11). Фактическое название этой должности не так важно. На практике важно, есть ли у этого человека поддержка руководства и бюджет на реализацию аналитической программы и продвижение корпоративной культуры на основе данных.

В приложении B приводится возможный вариант заявления о ви́дении компании в отношении данных. Заявление о ви́дении — это мотивирующее описание того, что компания стремится достичь в среднесрочной и долгосрочной перспективах. В данном случае компания стремится стать более ориентированной на данные в таких аспектах, как навыки работы с данными, повышение общей грамотности в вопросах работы с ними и формирование соответствующей корпоративной культуры. Обсу́дите этот документ с коллегами? Чего стремитесь достичь вы?

Самый верхний слой, в котором растворяются все остальные, — корпоративная культура, которая формирует все остальные слои и в равной степени сама формируется под их влиянием. Фактически управление на основе данных требует наличия в компании этих компонентов

и наиболее эффективных действий на каждом из этих уровней. Например, наличие в компании HiPPO может препятствовать объективному принятию решений на основе фактов. Политические игры и разобщенность данных негативно сказываются на открытости и сотрудничестве в рамках корпоративной культуры.

Многие компании прикладывают серьезные усилия, чтобы развить управление на основе данных. К сожалению, претворять в жизнь любые изменения, а особенно изменения культуры, крайне сложно. Шансы на развитие в компании успешной корпоративной культуры, основанной на данных, обычно выше, если начать заниматься этим как можно раньше, фактически создавая новую культуру, а не меняя ее. Это был один из мотивирующих факторов при написании этой книги. Я надеялся, что молодым компаниям, которые стремятся к управлению на основе данных и у которых еще впереди этап роста и привлечения новых сотрудников, это поможет стать более успешными. По результатам опроса, в котором приняли участие 368 стартапов[1], 3,26% респондентов заявили, что у них реализовано управление на основе данных: «С самого основания компании данные — часть нашей культуры». По словам еще 44% опрошенных, они «добились значительных улучшений и продолжают работать в направлении развития управления на основе данных». Это можно сравнить с изучением иностранного языка: многие успешно справляются с этой задачей во взрослом возрасте, но в детстве и юности учить иностранный язык бывает легче.

Еще один вопрос, который меня заинтересовал, — имеют ли некоторые онлайн-сервисы предрасположенность к управлению на основе данных, просто потому что они созданы вокруг продукта на основе данных. Возьмем, например, сайт знакомств, такой как OKCupid, рекомендательный сервис в области музыки Pandora или рекомендательный сервис в области контента Prismatic. Обязательно ли в подобных компаниях будет реализовано управление на основе данных в силу того, что их деятельность связана с данными и алгоритмами? Это вероятно, но не обязательно. Вполне возможно, что у таких компаний может быть ключевой продукт на основе данных, который развивается по принципам управления на основе данных, но, например, маркетинговые стратегии или привлечение клиентов подчиняются HiPPO.

Вероятно, здесь может иметь место явление, которое в популяционной генетике носит название «эффект основателя»[2], а в социальных

---

[1] Geckoboard and Econsultancy. Data Driven Culture: A global survey on the state of data driven culture in startups, 2013. URL: https://econsultancy.com/reports/data-driven-culture.

[2] URL: https://en.wikipedia.org/wiki/Founder_effect.

науках — «эффект колеи»[1]. Если в команде, которая сформировалась на старте проекта, высокая пропорция технических специалистов и специалистов по работе с данными, которые убеждены в необходимости применения аналитических инструментов и A/B-тестирования, это может повлиять на формирование соответствующей корпоративной культуры и задать тон в том, каких сотрудников компания будет нанимать в дальнейшем. Очевидно одно: в любой компании можно внедрить управление на основе данных. При конкуренции в области аналитики нет ограничений по сфере деятельности.

На протяжении всей книги я намеренно не делал акцента на технологиях. Не потому что это неважно, а потому что, по моему мнению, корпоративная культура в итоге — более весомый фактор. Позвольте объяснить мою точку зрения. Представьте, что в компанию приходит специалист по работе с данными и предлагает новейшие и самые эффективные инструменты (Spark, D3, R, библиотека Scikit-Learn и так далее). Если в корпоративной культуре компании не принято активно работать с данными, например там не проводят A/B-тестирование, а полагаются на мнение и опыт экспертов (HiPPO), работа специалиста по данным вряд ли окажет существенное влияние. Вероятно, он вскоре просто разочаруется и покинет компанию. А теперь представьте обратную ситуацию: в компании развита корпоративная культура на основе данных, но нет необходимых инструментов и технологий. Возможно, в компании ведутся основные реляционные базы данных, но до настоящего момента не возникала потребность в графовой базе данных или в кластере Hadoop. В подобных условиях у специалиста по работе с данными больше шансов получить финансирование и поддержку на разработку или приобретение любых инструментов, которые окажут влияние на эффективность деятельности компании. Иными словами, наличие правильных инструментов способно оказать огромное влияние. Но отсутствие правильной культуры или хотя бы стремления создать правильную культуру сведет на нет все усилия.

ВНИМАНИЕ: ВЗЛЕТ И ПАДЕНИЕ КОМПАНИИ TESCO

Tesco — британская транснациональная корпорация, крупнейшая розничная сеть в Великобритании и крупнейший работодатель в частном секторе (330 тыс. сотрудников). Ее называли эталоном компании с управлением на основе данных, конкурентное преимущество которой определяла ее аналитика.

[1] URL: https://en.wikipedia.org/wiki/Path_dependence.

В 1995 году компания запустила программу лояльности Clubcard. Это позволило аналитикам собрать данные о покупателях и поощрять их, таргетировав купоны. Благодаря более четкому таргетированию уровень погашения купонов вырос с 3 до 70%[1]. А за счет более точного сегментирования целевой аудитории компании удалось разработать и вывести на рынок новые продукты в верхнем ценовом сегменте (Tesco Finest), для тех, кто заботится о здоровье (Tesco Healthy Living), а также для тех, кому важно соотношение «цена/качество» (Tesco Value). В 1999 году объем их рассылки в разных сегментах составил 145 тыс. единиц.

Это был настоящий успех. Рыночная доля компании взлетела почти на 30%, Tesco стала крупнейшей розничной сетью в Великобритании. Сегодня у компании 16 млн активных участников программы лояльности и подробная информация о двух третях всех потребительских корзин. Покупатели получили более 1,5 млрд долл. в виде сэкономленных средств от использования баллов по программе лояльности. Компания выводила на рынок новые продукты специально для привлечения конкретных сегментов аудитории, например молодых родителей, и разрабатывала прогнозные модели, учитывавшие фактор погоды, для оптимизации цепочки поставок, что обеспечило экономию в объеме 150 млн долл. Компания занялась торговлей через интернет, предложив всем клиентам подписаться на программу лояльности Clubcard, и банковским делом. Сегодня Tesco вышла далеко за границы розничной торговли. По словам Майкла Шрейджа, «за исключением Amazon, ни одна глобальная розничная сеть не продемонстрировала более эффективного подхода, ориентированного на данные, касающиеся лояльности потребителей и их поведения»[2].

Аналитическим локомотивом за этим успехом был стартап Dunnhumby, в котором Tesco впоследствии выкупила контрольный пакет акций. Лорд Маклорин, бывший на тот момент председателем совета директоров компании, заявил супружеской чете основателей Dunnhumby: «Меня в этой ситуации пугает то, что спустя три месяца вы узнали о моих покупателях больше, чем я за 30 лет». Dunnhumby назвали «одной из жемчужин в короне Tesco».

Как дела у Tesco сегодня? Ее акции торгуются на самой низкой отметке за последние 11 лет. Компания потеряла 2,7 млрд долл. из-за неудачной попытки выйти на рынок США с сетью Fresh & Easy и объявила об убытке в объеме 9,6 млрд долл. за 2014 налоговый год.

---

[1] Patil R. Supermarket Tesco pioneers Big Data, Dataconomy, February 5, 2014. URL: http://dataconomy.com/2014/02/tesco-pioneers-big-data/.

[2] Schrage M. Tesco's Downfall Is a Warning to Data-Driven Retailers, Harvard Busines Review, October 28, 2014. URL: https://hbr.org/2014/10/tescos-downfall-is-a-warning-to-data-driven-retailers.

> Председатель совета директоров с позором покинул свой пост, после того как попытался завысить показатель прибыли на 400 млн долл. Компания сократила почти 9 тыс. рабочих мест и закрыла 43 магазина и их офисы. «С Tesco я допустил огромную ошибку», — признался Уоррен Баффет. Более того, Dunnhumby, чья программа лояльности Clubcard обходится в 750 млн долл. ежегодно (цена, при которой положительная рентабельность крайне маловероятна), выставлена на продажу за 3 млрд долл.
>
> Сложно выделить одну причину этого падения. Высокие показатели прибыли не помогли. Конкуренты разработали собственные программы лояльности, большинство из которых проще, а простота всегда привлекает! Вместо абстрактных «баллов» они предлагают своим клиентам более материальные бонусы, например газету или, что актуально для британцев, чашку чая[1].

К сожалению, управление на основе данных, и даже качественное управление на основе данных, не гарантирует успеха, а тем более устойчивого успеха. Во-первых, большинство успешных стратегий могут быть скопированы конкурентами, которые не преминут воспользоваться удачным опытом. Во-вторых, у руля компании все-таки стоит топ-менеджмент. И если руководство формулирует неверное ви́дение или стратегию для компании, даже решения, принятые на основе данных и поддерживающие эту стратегию, в итоге приведут к кораблекрушению. История Tesco, которую мы рассказали, — один из подобных примеров.

При этом на протяжении всей книги я приводил результаты разных исследований, свидетельствующие, что управление на основе данных окупается. Компаниям удается принимать решения быстрее и эффективнее и быстрее внедрять инновации. Компании, проводящие больше тестов, не только знают, когда что-то сработало, но и, скорее всего, знают, почему это произошло. Компании отличаются более высоким уровнем открытости, и любой сотрудник может внести свой вклад и увидеть, как это отразится на эффективности компании.

---

[1] Ruddick G. Clubcard built the Tesco of today, but it could be time to ditch it, The Telegraph, January 16, 2014. URL: http://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/10577685/Clubcard-built-the-Tesco-of-today-but-it-could-be-time-to-ditch-it.html.

Дополнительная литература

Аналитика

Aiken P. and Gorman M. The Case for the Chief Data Officer (New York: Morgan Kaufmann, 2013).

Davenport T. H. and Harris J. G. Analytics at Work (Boston: Harvard Business Press, 2007).

Davenport T. H., Harris J. G. and Morison R. Competing on Analytics (Boston: Harvard Business Press, 2010)[1].

Eckerson W. Secrets of Analytical Leaders: Insights from Information Insiders (Denville, NJ: Technics Publications, 2012).

Анализ данных

O'Neil C. and Schutt R. Doing Data Science (Sebastopol, CA: O'Reilly, 2014).

Shron M. Thinking With Data (Sebastopol, CA: O'Reilly, 2014).

Siegel E. Predictive Analytics (Hoboken: John Wiley & Sons, 2013)[2].

Silver N. The Signal and the Noise (New York: Penguin Press, 2012)[3].

Принятие решений

Kahneman D. 2011. Thinking, Fast and Slow. Farrar, Straus & Giroux, New York. Data Visualization[4].

---

[1] Издана на русском языке: Дэвенпорт Т., Харрис Дж. Аналитика как конкурентное преимущество. Новая наука побеждать. М. : BestBusinessBooks, 2010.

[2] Издана на русском языке: Сигель Э. Просчитать будущее. Кто кликнет, купит, соврет или умрет. М. : Альпина Паблишер, 2014.

[3] Издана на русском языке: Сильвер Н. Сигнал и шум. Почему одни прогнозы сбываются, а другие — нет. М. : Азбука-Аттикус: КоЛибри, 2000.

[4] Издана на русском языке: Канеман Д. Думай медленно… Решай быстро. М. : АСТ, 2016.

Визуализация данных

Few S. Now You See It (Oakland: Analytics Press, 2009).

Few S. Show Me the Numbers: Designing Tables and Graphs to Enlighten (Oakland: Analytics Press, 2012).

Tufte E. R. Envisioning Information (Cheshire, CT: Graphics Press, 1990).

Tufte E. R. Visual Explanations (Cheshire, CT: Graphics Press, 1997).

Tufte E. R. The Visual Display of Quantitative Information (Cheshire, CT: Graphics Press, 2001).

Wong D. M. The Wall Street Journal Guide To Information Graphics (New York: W. W. Norton & Company, 2010).

A/B-тестирование

Siroker D. and Koomen P. A/B Testing (Hoboken: John Wiley & Sons, 2013).

ПРИЛОЖЕНИЕ А

О необоснованной эффективности данных: почему больше данных лучше?

Данное приложение воспроизводится (с небольшими изменениями и исправлениями) на основе публикации в авторском блоге[1]*. Заголовок публикации сохранен.*

В научной работе The Unreasonable Effectiveness of Data («Необоснованная эффективность данных»)[2] авторы, все сотрудники компании Google, утверждают, что происходит интересная вещь, когда массивы данных попадают в вычислительную инфраструктуру (web scale[3]):

> Простые модели на основе большого объема данных значительно выигрывают у более сложных моделей на основе меньшего объема данных.

[1] URL: http://www.p-value.info/2012/12/on-unreasonable-effectiveness-of-data.html.

[2] Halevy A., Norvig P. and Pereira F. The Unreasonable Effectiveness of Data. Intelligent Systems, IEEE 24, no. 2 (2009): 8–12.

[3] Web scale — так аналитики Gartner определили термин, описывающий новый подход к вычислениям, разработанный и опробованный на практике такими облачными провайдерами, как Google, Amazon, Rackspace, Netflix, Facebook и другими. Фактически это инновационная методология построения дата-центров и программной архитектуры, совокупно объединяющей такие разные концепции, как масштабируемость, интегрируемость, устойчивость к сбоям, специализация и пр. *Прим. науч. ред.*

В этой научной работе и более подробной лекции, прочитанной Норвигом[1], авторы демонстрируют: когда размер обучающей выборки доходит до сотен миллионов или триллионов примеров, очень простые модели способны быть эффективнее более сложных, основанных на тщательно разработанных онтологиях, но на меньшем объеме данных. К сожалению, авторы практически не предоставляют объяснений, *почему* больше данных лучше. В этом приложении я хочу попытаться найти ответ на этот вопрос.

Мое предположение состоит в том, что существует несколько типов проблем и причин, почему больше данных лучше.

Проблемы типа «ближайший сосед»

Первый тип проблем можно условно назвать «*ближайший сосед*». Халеви и др. приводят пример:

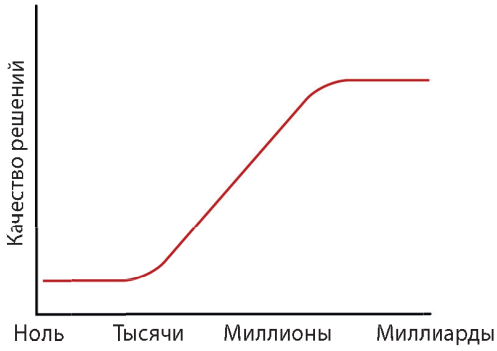
> Джеймс Хейс и Алексей Эфрос занялись задачей дополнения сцены: они решили удалить фрагмент изображения (портящий вид автомобиль или бывшего супруга) и заменить фон путем добавления пикселей, взятых из большого набора других фотографий[2].

Рисунок 1 Хейса и Эфроса

[1] URL: https://www.youtube.com/watch?v=yvDCzhbjYWs.

[2] Hays J. and Efros A. A. Scene Completion Using Millions of Photographs. Proceedings of ACM SIGGRAPH 2007, San Diego, CA, August, 5–9, 2007, pp. 1–7. URL: http://graphics.cs.cmu.edu/projects/scene-completion/scene-completion.pdf.

Норвиг изобразил следующую зависимость:



и описал ее как «порог данных», при котором результаты из очень плохих стали очень хорошими.

Я не уверен, что существует какая-то пороговая величина или что-то напоминающее фазовый переход. Скорее, мне кажется, суть проблемы заключается в поиске ближайшего соответствия. Чем больше данных, тем ближе может быть соответствие.

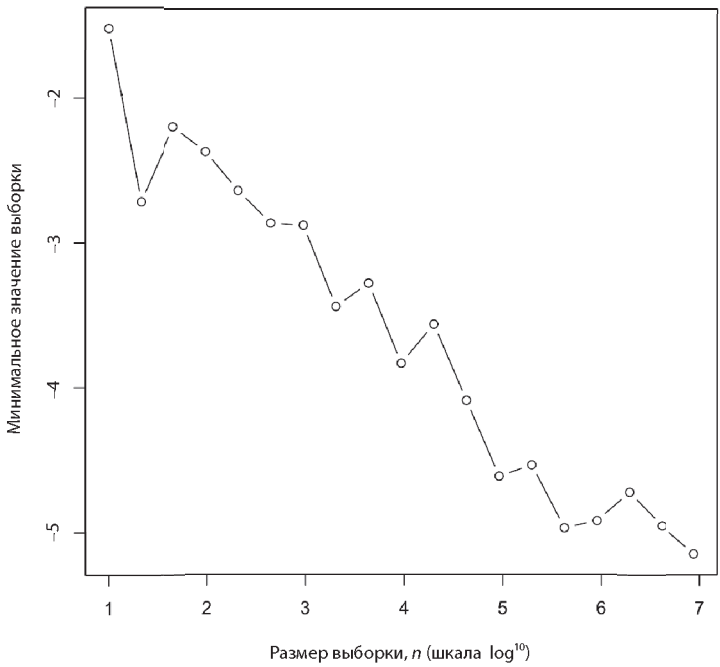
Хейс и Эфрос отмечают:

> Результаты наших первых экспериментов с GIST-дескриптором по базе данных из 10 тыс. изображений крайне нас разочаровали. Тем не менее при увеличении размера набора данных до 2 млн единиц произошел качественный скачок… Независимо от нас Торралба и др. [2007] наблюдали похожий эффект с базой данных размером до 70 млн небольших (32×32) изображений… Для успеха нашего метода требуется большой объем данных. Мы наблюдали существенное улучшение, когда перешли от 10 тыс. к 2 млн изображений.

Размеры двух этих наборов данных различаются слишком сильно, а «качественный скачок» — это не то же самое, что порог (буквально фазовый переход).

Увеличение объема данных может значительно повлиять на показатели из-за простых эффектов. Например, рассмотрим выборку размера n в стандартном нормальном распределении. Как изменяется в зависимости от значения n минимальное значение этой выборки? Создадим выборки разных размеров и вычислим минимальное значение с помощью следующего кода R:

```
x<-seq(1,7,0.5)
y<-vector(mode="numeric",length=length(x))
for (i in 1:length(x)){ y[i] <- min(rnorm(10^(x[i]))) }
plot(x,y,xlab="Sample size, n (log10 scale)",
ylab="Minimum value of sample",type="b")
```



Минимум уменьшается лог-линейно. Это случай экстремума с позиции неограниченного хвоста. Возможно, более подходящей здесь для проблемы минимизации, такой как подбор соответствия, будет нижняя граница — идеальное соответствие для всех целей. Например, возможно, кто-то еще, стоя на том же самом месте, сделал фотографию того же самого вида, но без предмета, портящего фотографию.

Думаю, именно это происходит на графике Норвига. При определенном размере выборки мы нашли очень хорошее соответствие, и увеличение размера выборки уже не может улучшить результат.

Подведем итог: для проблемы минимизации типа «ближайший сосед» с неотрицательной функцией расстояния (что означает, что нижняя граница функции ошибки обучения (cost function) равна нулю) функция расстояния в среднем будет монотонно убывать с размером выборки или данных.

Проблемы относительной частотности

Второй тип — это проблемы *относительной частотности*. Именно на них сосредоточились Халеви и др. Норвиг приводит несколько примеров. При сегментировании задача заключается в разделении исходного текста, например такого как «cheapdealsandstuff.com», на наиболее вероятные последовательности слов. Эти исходные варианты достаточно короткие, чтобы с ними можно было работать непосредственно с позиции возможного их разделения, но для каждого получившегося отдельного слова нужно оценить вероятность его существования. Самое простое предположение — о независимости среди слов. Таким образом, если Pr (*w)* — это вероятность слова *w,* то, имея некоторый набор данных, можно вычислить, например:

```
Pr(che,apdeals,andstuff) = Pr(che) . Pr(apdeals) . Pr(andstuff).
...
Pr(cheap,deals,and,stuff) = Pr(cheap) . Pr(deals) . Pr(and) .
Pr(stuff).
```

Конечно, также можно использовать *n*-граммы (например, биграммы): `Pr("cheap deals") × Pr("and stuff").`

Второй пример, который привел Норвиг, касался проверки орфографии. В этом случае можно взять слово, содержащее ошибку, и вычислить вероятность возможных вариантов, чтобы предложить наиболее вероятную форму.

В обоих случаях требуется набор данных, содержащий как характерные, так и нехарактерные слова и фразы. Кроме того, необходим показатель встречаемости этих фраз для вычисления относительной частотности. Чем больше и понятнее будет набор данных, тем лучше. Думаю, здесь наблюдаются два статистических явления.

— Чем больше корпус данных, тем выше качество оценки относительной частотности. Это закон больших чисел[1].

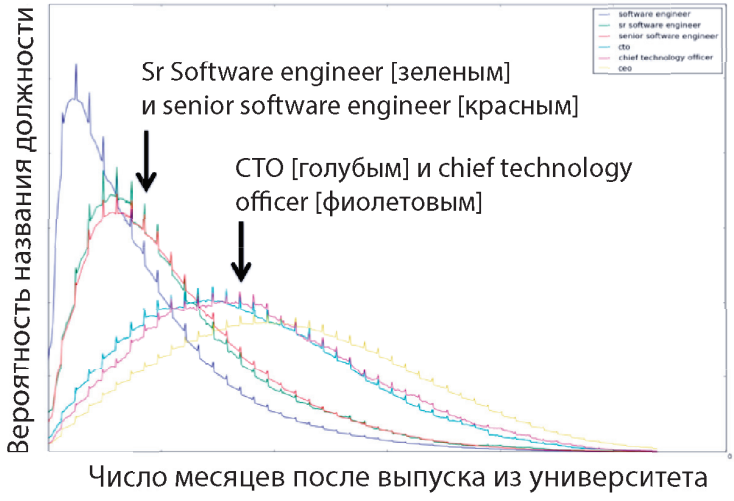
— Чем больше корпус данных, тем выше вероятность попадания в него нехарактерных фраз («длинного хвоста»). Это неограниченный эффект. Чем больше индексируется интернет, тем больше новых фраз будет появляться. Проблема осложняется тем, что распределение слов в английском языке — это степенной закон. (См. Zipf, G. The Psycho-Biology of Language. Houghton Mifflin, Boston, MA, 1935.) Это означает наличие особенно длинного хвоста. Сле-

[1] URL: https://en.wikipedia.org/wiki/Law_of_large_numbers.

довательно, особенно крупные выборки должны содержать эти редкие фразы.

Проблемы оценки одномерного распределения

К третьему типу относятся проблемы оценки одномерного распределения. Недавно я слушал лекцию[1] Питера Скомороха из компании LinkedIn[2]. Он показал распределение вероятности названия должности сотрудника, занимающегося разработкой программного обеспечения, в зависимости от числа месяцев, прошедших после его выпуска из университета. Согласно данным, распределения «Sr Software engineer» и «senior software engineer» (старший инженер-разработчик программного обеспечения) почти идентичны, что можно было ожидать, учитывая их синонимичность. Аналогичная картина и с распределениями «CTO» и «Chief Technology Officer». Это интересный способ определения синонимов и исключения повторов, вместо того чтобы поддерживать длинный основной список акронимов и аббревиатур. Это возможно только благодаря объему данных: при нем распределение, которое делают авторы, — надежное и предположительно близкое к истинному лежащему в основе распределению населения.


Вероятность названия должности
Sr Software engineer [зеленым]
и senior software engineer [красным]
CTO [голубым] и chief technology
officer [фиолетовым]
Число месяцев после выпуска из университета

Источник: Питер Скоморох. Воспроизводится с разрешения

[1] Skomoroch P. Developing Data Products, December 5, 2012. URL: https://www.slideshare.net/pskomoroch/developing-data-products.

[2] Analytics Talk: Peter Skomoroch, December 13, 2012. URL: https://www.airbnb.ru/meetups/ejs83rxek-analytics-talk-peter-skomoroch.

Проблемы многофакторности

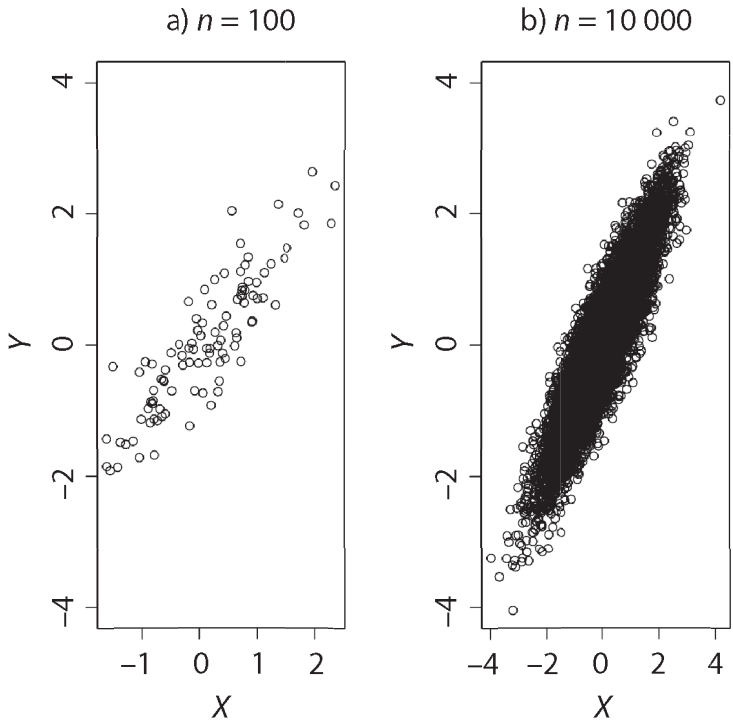
Четвертый тип проблем — проблемы *многофакторности,* или *корреляционные,* при которых мы стремимся оценить взаимоотношения между переменными. Это может быть оценка взаимоотношений $y = f(x)$ или, возможно, оценка совместной плотности распределения многих переменных. Это можно использовать для разрешения лексической многозначности (например, когда в документе встречается слово pike, обозначает ли оно «щуку» или «пику») или для составления «справочника» взаимосвязанных характеристик или концепций для конкретной лексической единицы (например, с понятием «компания» связаны такие понятия, как «генеральный директор», «главный офис», «ИНН» и так далее).

В данном случае нас интересуют корреляции между словами или фразами. Проблема в том, что документы в сети отличаются высокой размерностью, и, принимаясь за решение подобных проблем, мы попадаем под действие «проклятия размерности»[1], когда данные становятся очень рассеянными.

Таким образом, один из эффектов более крупной выборки заключается в повышении плотности данных в статистическом пространстве. Опять-таки, в случае с более крупными выборками есть возможность более точно оценить показатели, такие как показатели положения (среднее значение, медиана и другие показатели центра распределения). Кроме того, можно более точно оценить совместные плотности распределения (PDFs). Следующая диаграмма рассеяния представляет собой простой пример, составленный на основе этого кода:

```
par(mfrow=c(1,2))
plot(mvrnorm(100, mu = c(0, 0),
Sigma = matrix(c(1, .9, .9, 1), 2)),xlab="X",ylab="Y",
ylim=c(-4,4))
title("n = 100")
plot(mvrnorm(10000, mu = c(0, 0),
Sigma = matrix(c(1, .9, .9, 1), 2)),xlab="X",ylab="Y",
ylim=c(-4,4))
title("n = 10000")
```

[1] URL: https://en.wikipedia.org/wiki/Curse_of_dimensionality.



Слева использовалась маленькая выборка. Диаграмму легко интерпретировать как линейную. Справа, где размер выборки был больше, более очевидно настоящее двумерное нормальное распределение. Конечно, это банальный пример. Суть в том, что для более высоких размерностей требуется значительно более серьезный размер выборки, чтобы также оценить совместные плотности распределения.

Конечно, это весьма поверхностный ответ на вопрос, почему больше данных лучше. Предпочтительно использовать качественные данные. Однако во многих компаниях, таких как Google, Twitter, LinkedIn и Facebook, где контент создается пользователями, нередко тексты, созданные в свободной форме, касаются самых разных областей (поэтому глубокая очистка данных и использование онтологий просто нерациональны), в итоге мы видим, что «информационный шум» компенсируется очень большим объемом данных. В итоге все выравнивается, и в случае проблем «ближайшего соседа» решение всегда будет лучше.

ПРИЛОЖЕНИЕ В

Заявление о видении

Это приложение может стать стартовой точкой для формирования заявления о ви́дении — мотивирующего описания того, чего компания стремится достичь в среднесрочной и долгосрочной перспективах, чтобы стать более ориентированной на данные. Суть в том, чтобы выделить цель компании, объединить всех участвующих лиц и стимулировать обсуждение того, как добиться целей компании. Каждая компания индивидуальна, скорректируйте этот документ так, чтобы он отражал ви́дение вашей компании.

В процветающей компании с управлением на основе данных [название компании] присутствует следующее.

Сильное руководство на основе данных

— Руководители активно продвигают данные как стратегический актив, который должен максимально использоваться для оказания влияния на все уровни деятельности компании.

— Руководители понимают потребности бизнеса и поддерживают его развитие. Руководители поддерживают специалистов аналитического подразделения: обеспечивают им четкий карьерный путь, стимулируют работать максимально эффективно и получать удовольствие от работы.

— Менеджеры опираются на аналитические выводы для принятия информированных решений. В целом в компании использование данных и аналитики глубоко укоренилось в наших рабочих процессах и процессе принятия решений.

Открытая культура, построенная на доверии

— Существует централизованный набор связанных источников данных без барьеров.

— У бизнес-подразделений сформирована концепция владения знаниями, сотрудники активно управляют качеством данных из своих источников.

— Обеспечен широкий доступ к данным.

  а) У каждого сотрудника, которому требуется доступ к данным для выполнения своих функциональных обязанностей, есть этот доступ.

  б) У каждого сотрудника есть доступ только к тем данным, которые необходимы ему для выполнения своих функциональных обязанностей. Работа с персональными данными, например с информацией о клиентах или рекомендациями, ведется особенно внимательно: доступ к таким данным существенно ограничен, данные обезличены и закодированы.

  в) Каждый сотрудник компании может легко получить целостное представление обо всей деятельности компании благодаря доступным и понятным дашбордам, отчетам и аналитическим выводам. Системы раннего предупреждения оборудованы необходимыми инструментами и надежны.

— Специалисты по аналитике активно взаимодействуют со всеми подразделениями компании и помогают оценить идеи и проверить их объективность.

Самодостаточная система аналитики

— Процесс работы со стандартной отчетностью полностью автоматизирован. Бóльшую часть рабочего времени специалисты по аналитике тратят на проведение специализированного анализа, поиск источников данных и прогнозное моделирование и оптимизацию.

— С помощью инструментов бизнес-аналитики осуществляется стандартный поиск данных, а интерфейс SQL поддерживает все остальные специализированные запросы.

Широкая функциональная грамотность при работе с данными

— Все сотрудники аналитического подразделения обладают основными аналитическими и статистическими навыками в соответствии с их должностью.

— Все лица, принимающие решения, в том числе топ-менеджмент компании, обладают функциональной грамотностью при

работе с данными, могут интерпретировать статистические выводы и оценить качество проведения экспериментов.

— Существуют широкие возможности для обмена знаниями, обучения и совершенствования своих навыков благодаря участию в семинарах и курсах, чтению специальной литературы и принципам наставничества.

Объективная культура, в которой сначала устанавливаются цели

— Существует четко сформулированное, разделяемое всеми сотрудниками, доступное ви́дение, к каким целям стремится компания. Ее стратегия, действия и тактика стимулируются прозрачной и часто упоминаемой системой ключевых показателей эффективности деятельности.

Культура, в которой задают вопросы

— В компании сформирована уважительная среда, в которой приветствуются конструктивные обсуждения, и каждый сотрудник может задать вопрос другим относительно их данных, предположений и аналитической интерпретации.

— «У вас есть данные, подтверждающие это?» — никто не должен бояться задавать этот вопрос, и все должны быть готовы на него ответить.

Культура, в которой проводятся тестирования

— Все рациональные идеи проходят тестирование (как онлайн, так и офлайн): сбор данных, изучение, повторение. Объективные эксперименты — норма.

Ценность

Конечно, вы должны обосновать, *почему* сотрудники должны принять это ви́дение.

Финансы

При прочих равных условиях эффективность деятельности компании с управлением на основе данных на 5–6% выше, чем у других, не опирающихся на данные. Кроме того, у такой компании более эффективное использование ресурсов, выше рентабельность собственных средств и рыночная ценность.

Рентабельность аналитики составляет 13,01 долл. на каждый вложенный доллар.

Руководство на основе данных

Централизованный подход к аналитической работе и поддержка со стороны руководства повышают у специалистов по аналитике степень удовлетворенности своей работой и снижают вероятность, что они захотят покинуть компанию.

Самодостаточность

Если сотрудники разных подразделений обладают навыками статистической работы и планирования экспериментов и хотя бы один сотрудник у них умеет работать с SQL, они будут более самодостаточными, независимыми, с более высокой скоростью реакции и масштабом деятельности.

Проведение тестов

Сотрудники принимают решения на основе качественных и количественных данных, полученных от настоящих покупателей. Им не приходится догадываться, как покупатели могут отреагировать на новую функцию.

Имея возможность проводить тестирования и интерпретировать их результаты, компания может быстрее внедрять инновации. За месяц сотрудники могут протестировать десятки или сотни идей по оптимизации сайта.

Реализация

Наконец, вам необходимо согласовать фактический план действий, *как* вы собираетесь реализовывать это ви́дение. Чего вы ожидаете от коллег?

Руководство на основе данных

Согласуйте матрицу аналитических компетенций.

Поднимите планку качества для новых и действующих специалистов по аналитике. Стимулируйте действующих аналитиков развивать свои навыки.

Открытость и доверие

Займите активную позицию в отношении качества данных. Разработайте систему обзора, оповещений и других способов контроля для отслеживания объема данных, их качества и возможных проблем.

Самодостаточность

Изучите SQL. Команды всех бизнес-подразделений должны стать более самодостаточными и уметь проводить более специализированные исследования.

Умение работать с данными

Все менеджеры должны уметь работать со статистикой.

Объективность и постановка целей

Свяжите все проекты с главными стратегическими целями компании. Каждому сотруднику должно быть ясно, почему в компании осуществляется или не осуществляется тот или иной проект и как расставлены приоритеты.

По возможности оперируйте конкретными цифрами, например ROI.

Для любого компонента корпоративной культуры, который вы захотите внедрить в своей компании, вам потребуется ответить на вопросы *что, почему* и *как*.

Благодарности

Эта книга стала результатом совместного вклада в виде идей и помощи от коллег и экспертов. Я хочу выразить благодарность за чрезвычайно полезные советы, рекомендации и поддержку очень многим людям. Вот они: Эндрю Абел, Питер Айкен, Трейси Эллисон Олтман, Самарет Баскар, Лон Биндер, Нейл Блументаль, Йозеф Боренштайн, Льюис Брум, Трей Кози, Брайн д'Алессандро, Грег Элин, Саманта Эверитт, Марио Фариа, Стивен Фью, Том Фишбурн, Эндрю Фрэнсис Фриман, Дейв Джилбо, Кристина Ким, Ник Ким, Анджали Кумар, Грег Линден, Джейсон Гоуэнс, Себастьян Гутьеррес, Дуг Лейни, Шон Лисен, Дуг Мак, Патрик Махони, Крис Малиуот, Микайла Маркрич, Линн Массимо, Санья Матур, Мириа Мейер, Джули-Дженнифер Нгуен, Скотт Поли, Джефф Поттер, Мэтт Риццо, Макс Шрон, Анна Смит, Неллвин Томас, Дэниел Танкеланг, Джеймс Валландингхэм, Сатиш Ведантам, Дэниел Уайт и Дэн Вудс.

Кроме того, я благодарю всех своих коллег из Warby Parker, оказавших мне серьезную поддержку.

Мои искренние извинения всем, кого я ненамеренно не упомянул.

Особая моя благодарность Дэниелу Минтцу, Джули Стил, Дэну Вудсу, Лону Биндеру и Джун Эндрюс, выступившим в качестве технических редакторов и предложивших обоснованные и ценные комментарии, которые помогли мне значительно улучшить книгу.

Спасибо организаторам Data Driven Business, особенно Антанине Капчонава, и участникам форума Chief Data Officer Executive Forum, состоявшегося 12 ноября 2014 года в Нью-Йорке. Джеймс Валландингхэм внес изменения в рис. 4.1 специально для этой книги. Спасибо, Джим!

Хочу поблагодарить Себастьяна Гутьерреса за содержательную беседу и разрешение использовать некоторые примеры из его отличного курса по визуализации данных.

Я не могу обойти вниманием поддержку своих друзей и семьи, особенно моей жены Алексии, которая в шутку называла себя «книжной вдовой», а также моей мамы, которая поддерживает меня на протяжении всей жизни.

Наконец, невозможно не выразить благодарность всей великолепной команде издательства O'Reilly, особенно редактору книги Тиму Макговерну. Я признателен за проделанную работу Майку Лукидесу, Бену Лорика, Мари Богуро и производственной команде: Коллину Лобнеру, Люси Хаскинс, Дэвиду Футато, Киму Коферу, Элли Волькхаузен, Аманде Керси и Ребеке Демарест.

Об авторе

Карл Андерсон — директор направления по работе с данными компании Warby Parker в Нью-Йорке. Он отвечает за технические аспекты этого направления, поддерживает более широкую аналитическую структуру и развивает в компании корпоративную культуру на основе данных. До этого работал преимущественно в области применения вычислительных машин для решения научных задач в разных компаниях из таких сфер деятельности, как моделирование в здравоохранении, сжатие данных, робототехника, моделирование с применением исполнительных устройств. Имеет степень Ph.D. в области математической биологии, полученную в Университете Шеффилда, Великобритания.

Колофон

Птица, изображенная на обложке книги, это трехцветный спрео, или великолепный скворец (*Lamprotornis superbus*). Эта певчая птица семейства скворцовых обитает в восточной части Африканского континента от Эфиопии до Танзании.

Взрослые особи отличаются оперением очень красивого цвета: вверху блестящее черное, на затылке и плечах блестящее сине-зеленое. Шея, горло и грудь металлически-синего блестящего цвета. Полоса на груди и гузка белые, брюхо окрашено в красно-бурый цвет. Длина взрослых птиц составляет примерно 18 см, а размах крыльев до 40 см.

Птицы очень «социализированы» и общаются при помощи длинных призывных трелей. Живут обычно в больших стаях и часто совместно заботятся о потомстве. Их пища состоит в основном из насекомых, плодов и семян, но если предоставляется такая возможность, то могут назойливо выпрашивать корм в деревнях или городах.

Многие из представителей животного мира, которых издательство O'Reilly помещает на обложки, находятся на грани вымирания. Все они важны для нашей планеты. Узнать подробнее о том, как вы можете помочь, можно на сайте *animals.oreilly.com*.

Максимально полезные книги

Заходите в гости:
http://www.mann-ivanov-ferber.ru/

Наш блог:
http://blog.mann-ivanov-ferber.ru/

Мы в Facebook:
http://www.facebook.com/mifbooks

Мы ВКонтакте:
http://vk.com/mifbooks

Предложите нам книгу:
http://www.mann-ivanov-ferber.ru/about/predlojite-nam-knigu/

Ищем правильных коллег:
http://www.mann-ivanov-ferber.ru/about/job/

Научно-популярное издание

Карл **Андерсон**

Аналитическая культура
От сбора данных до бизнес-результатов

Главный редактор *Артем Степанов*
Ответственный редактор *Светлана Мотылькова*
Литературный редактор *Юлия Слуцкина*
Арт-директор *Алексей Богомолов*
Верстка обложки *Наталия Майкова*
Верстка *Екатерина Матусовская*
Корректоры *Мария Кантурова, Надежда Болотина*

O'REILLY®

Что нужно, чтобы внедрить в компании управление на основе данных? Отладить сбор «больших данных»? Собрать команду аналитиков? Да, но этого недостаточно. В первую очередь необходима конструктивная корпоративная культура. Это неписаный свод правил, которые затрагивают качество данных и обмен информацией, прием на работу аналитиков и их обучение, коммуникации, процессы принятия решений и многое другое.

Благодаря данному практическому руководству, в основу которого легли интервью с ведущими аналитиками известных компаний, вы поймете, как создать такую культуру. Карл Андерсон, директор по аналитике в компании Warby Parker, рассказывает, какие процессы следует внедрить на всех уровнях: от аналитиков и менеджеров до высшего руководства и совета директоров — и как это сделать.

Вы узнаете об аналитической цепочке ценности, которая поможет принимать правильные решения и достигать лучших бизнес-результатов.

— Начните с начала: узнайте, как собирать правильные данные.
— Найдите аналитиков, обладающих правильными навыками, и соберите их в команды.
— Изучите статистические методы и инструменты визуализации данных.
— Соберите и проанализируйте данные, соблюдая конфиденциальность и не забывая об этике.
— Выясните, как аналитики способствуют развитию культуры на основе данных.

ISBN 978-5-00100-781-4
9 785001 007814 >

Максимально полезные книги для детей на сайте **mann-ivanov-ferber.ru**

издательство
МАНН, ИВАНОВ И ФЕРБЕР

Like facebook.com/**mifbooks**
vk.com/**mifbooks**
instagram.com/**mifbooks**